ulm university universität
uulm

Differenzierte Messung kognitiver Belastung beim Lernen im Rahmen von Instruktionsdesignfragestellungen

kumulative Dissertation zur Erlangung des Doktorgrades
– *Dr.rer.nat.* –
der Fakultät für Ingenieurwissenschaften, Informatik und Psychologie der Universität Ulm

Melina Klepsch
aus Bregenz

2020

Abteilung Lehr-Lernforschung
Institut für Psychologie und Pädagogik
Fakultät für Ingenieurwissenschaften, Informatik und Psychologie
Universität Ulm

Amtierender Dekan: Prof. Dr.-Ing. Maurits Ortmanns
Gutachter: Prof. Dr. Tina Seufert, Prof. Dr. Joachim Wirth
Datum der Promotion: 04. September 2020

Mit den Steinen die mir heute in den Weg gelegt werden, baue ich später mein Traumhaus.

(Unbekannt)

Vorwort und Danksagung

All denen, die mich beim Zustandekommen dieser Arbeit im Großen wie im Kleinen unterstützt haben, möchte ich herzlich danken, denn jeder/jede einzelne Genannte hat während der Dauer meiner Promotionszeit bestimmt mal sehr viel Geduld mit mir haben müssen. Es gab Hochs und Tiefs, es gab produktive und unproduktive Zeiten, es gab anstrengende und erfreuliche Zeiten.

Mein Dank gilt meiner Betreuerin Prof. Dr. Tina Seufert für ihre Unterstützung und dafür, dass sie mir den Freiraum zur Verwirklichung meiner Vorstellungen ließ. Deine zahlreichen Anregungen haben mir immer wieder neue Blickwinkel eröffnet und den Fortgang dieser Arbeit gefördert. In den von dir eingeführten und mehrmals im Jahr stattfindenden Schreibwochen, die ich jedes mal sehr produktiv fand, habe ich große Teile dieser Dissertation fertigstellen können, ohne das Gefühl zu haben in dieser Zeit meine anderen Aufgaben zu vernachlässigen. Dafür ein herzliches Danke, behalte diese wertvolle Zeit bei, auch andere Promovierende werden dir dafür bestimmt danken. Danke auch dafür, dass deine Tür immer offen steht.

Mein Dank geht auch an alle wissenschaftlichen Mitarbeiter, Sekretärinnen und Hilfskräfte währende meine Promotionszeit in der Abteilung Lehr-Lernforschung. Sie alle haben dafür gesorgt, dass ich gerne an meiner Promotion gearbeitet haben, dass ich trotz vielfältiger Aufgaben immer wieder zu meiner Promotion zurückgefunden haben und mir die Zeit sehr positiv in Erinnerung bleiben wird. Insbesondere möchte ich hier meinen (ehemaligen) Kolleginnen Lisa Respondek, Nadja Müller und Janina Lehmann danken, die den ein oder anderen Knoten in meinem Kopf gelöst haben und geholfen haben das große Ganze aus immer wieder anderen Blickwinkeln zu betrachten. Danke auch an Rebecca Pientka die mir gerade gegen Ende meiner Promotionszeit den Rücken freigehalten hat.

Mein besonderer Dank gilt meinem Mann, Matthias Klepsch, der diese Dissertation wie niemand sonst durch seinen Glauben an mich während meiner gesamten Promotionsphase begleitet und alles erst möglich gemacht hat. Ein Dank auch an meine drei kleinen Rabauken, die all meine Freizeit für sich beanspruchen und so die kleinen Fenster konzentrierter wissenschaftlicher Arbeit noch produktiver erscheinen lassen.

Danke auch an meine Eltern, die jeden meiner Schritte begleiten und immer Unterstützend da sind wenn ich sie brauche.

Überblick relevanter Beiträge

Die vorliegende Dissertation ist eine kumulative Dissertation, die sich mit der Messung kognitiver Belastung im Rahmen von Instruktionsdesignfragestellungen beschäftigt. Diese Dissertation umfasst 9 Beiträge mit zusammen 12 verschiedenen Studien.

Beitrag I - Zeitschriftenartikel

Klepsch, M., Schmitz, F. & Seufert, T. (2017). Development and Validation of Two Instruments Measuring Intrinsic, Extraneous, and Germane Cognitive Load. *Frontiers in Psychology, 8*, 1997. doi: 10.3389/fpsyg.2017.01997

Beitrag II - Artikel in Tagungsband & Tagungsbeitrag

Pichler, M. & Seufert, T. (2011). Two strategies to measure Cognitive Load. In European Association for Research on Learning and Instruction (Hg.), *EARLI Conference 2011 Ëducation for a Global Networked Society": Book of Abstracts and Extended Summaries* (S. 928–929). University of Exeter.

Beitrag III - Tagungsbeitrag (Akzeptierte Einreichung eingefügt)

Klepsch, M. & Seufert, T. (9. April 2012). *Subjective Differentiated Measurement of Cognitive Load.* 5th International Cognitive Load Theory Conference, Tallahassee (USA).

Beitrag IV - Zeitschriftenartikel

Klepsch, M. & Seufert, T. (2020). Understanding instructional design effects by differentiated measurement of intrinsic, extraneous, and germane cognitive load. *Instructional Science, 48*(1), 45–77. doi: 10.1007/s11251-020-09502-9

Beitrag V - Tagungsbeitrag (Akzeptierte Einreichung eingefügt)

Klepsch, M., Kempter, I. & Seufert, T. (2013). Interaction of Worked Examples and Prompts: Impact on Performance and Cognitive Load. In European Association for Research on Learning and Instruction (Hrsg.), *15th Biennial EARLI Conference "Responsible Teaching and Sustainable Learning"*, München (DE).

Beitrag VI - Artikel in Tagungsband und Tagungsbeitrag

Klepsch, M., Westphal, J. & Seufert, T. (2013). Effekte von integriertem bzw. separiertem Text-Bild-Material in Abhängigkeit von serialistischem oder holistischem Lernstil., *14. Fachgruppentagung Pädagogische Psychologie*. Hildesheim (DE).

Beitrag VII - Artikel in Tagungsband und Tagungsbeitrag

Seufert, T., Klepsch, M. & Westphal, J. (2017). Adjusting Task Difficulty to Control Learner's Intrinsic Cognitive Load. In European Association for Research on Learning and Instruction (Hrsg.), *EARLI 2017 Ëducation in the crossroads of economy and politics - Role of research in the advancement of public good"*, Tempere (FIN).

Beitrag VIII - Zeitschriftenartikel und Tagungsbeitrag

Rogers, K., Röhlig, A., Weing, M., Gugenheimer, J., Könings, B., Klepsch, M., Schaub, F., Rukzio, E., Seufert, T. & Weber, M. (2014). P.I.A.N.O: Faster Piano Learning with Interactive Projection. In R. Dachselt, N. Graham, K. Hornbæk & M. Nacenta (Hrsg.), *Ninth ACM International Conference on Interactive Tabletops and Surfaces*, Dresden (DE). doi: 10.1145/2669485.2669514

Beitrag IX - Artikel in Tagungsband und Tagungsbeitrag

Klepsch, M., Könings, B., Weber, M. & Seufert, T. (25. August 2014). Fostering Piano Learning by dynamic mapping of notes. European Association for Research on Learning and Instruction SIG 2 (Hrsg.), *EARLI SIG 2 Meeting "Building bridges: Improving our understanding of learning from text and graphics by making the connection"*, Rotterdam (NE).

Zusammenfassung

Die Cognitive Load Theory (CLT) ist eine der bedeutendsten Theorien im Forschungsfeld der Lehr-Lernforschung. Die CLT nimmt an, dass unser Arbeitsgedächtnis der Flaschenhals beim Wissenserwerb ist, während das Langzeitgedächtnis in seiner Kapazität als unbegrenzt gilt. Die zentrale Aufgabe von Lehrenden und Instruktionsdesigner besteht daher darin, die kognitive Belastung, welche auf das Arbeitsgedächtnis eines Lernenden wirkt möglichst optimal zu verteilen. Dabei werden in der CLT klassisch drei Belastungsarten unterschieden, die Arbeitsgedächtnisressourcen beanspruchen können: (1) Der Intrinsic Cognitive Load (ICL) ergibt sich einerseits aus der Komplexität des Lerninhaltes, der über das Ausmaß an Elementinteraktivität definiert wird, und andererseits aus dem Vorwissen des Lernenden. (2) Der Extraneous Cognitive Load (ECL) entsteht durch die Gestaltung des Lernmaterials, eine suboptimale Gestaltung erhöht diese Belastung und beeinträchtigt ggf. den Lernerfolg. (3) Der Germane Cognitive Load (GCL) ist die dritte Quelle kognitiver Belastung, es ist die lernbezogene und lernförderliche Belastung und entsteht durch Konstruktion von Schemata und aktiver Auseinandersetzung mit dem Lerninhalt.

Neben theoretischen Diskussionen, über die konzeptionellen Teile der Theorie, ist die größte Herausforderung, das Messen der kognitiven Belastung beim Lernen. In dieser Promotion wird die Entwicklung und Validierung einer differenzierten Messung der verschiedenen Arten kognitiver Belastung vorgestellt und diskutiert. Im Rahmen der Promotion wurden theoriegeleitet drei subjektive Skalen zur Messung der unterschiedlichen Arten kognitiver Belastung entwickelt. Diese wurden mit einem Expertenrating vergleichen, sowie deren Faktorstruktur näher betrachtet. Dabei stellte sich heraus, dass eine differenzierte Messung der verschiedenen Arten kognitiver Belastung durchaus möglich ist.

In einem weiteren Schritt wurde der Fragebogen in weiteren Studien eingesetzt, um zu prüfen ob sich die theoretischen Annahmen im Lernmaterial der Studien auch in den Skalen widerspiegeln. Die Studien konnten erfolgreich zeigen, dass eine differenzierte subjektive Messung zwischen verschiedenen Instruktionen Veränderungen differenzieren kann und die theoretischen Annahmen abbilden kann. Auch im praktischen Einsatz konnte der Fragebogen erfolgreich eingesetzt werden.

Abschließend werden noch theoretische, methodische und praktische Implikationen dargelegt und diskutiert. Dabei werden Kritikpunkte an der Theorie wie an den Messmethoden aufgegriffen und in einem weiteren Schritt jeweils aufgezeigt, wie der entwickelte Fragebogen helfen kann die Kritikpunkte empirisch in weiterer Forschung anzugehen.

Inhaltsverzeichnis

Teil I.

Motivation, theoretische Hintergründe und verwandte Arbeiten

Wenn du Heute aufgibst, wirst du nie wissen ob du es Morgen geschafft hättest!

(unbekannt)

1. Motivation und Ziele dieser Arbeit

Digitalisierung ist im Lernkontext längst Alltag - im Klassenzimmer, im Hörsaal, im Betrieb, ... wir lernen mit Laptop, Tablet und am Smartphone, manchmal sogar in virtuellen Realitäten. Wir lernen im formalen Rahmen, im non-formalen Rahmen und informell. Wir lernen weil wir müssen und/oder weil wir wollen.

Immer öfter nutzen wir durch die Digitalisierung multimediale und multiple Repräsentationen zum Lernen. Doch der Einsatz digitaler Medien ist eigentlich nur ein Mittel, das Ziel ist nach wie vor dasselbe: Wir wollen unser Wissen und unsere Fähigkeiten erweitern.

Dies geschieht durch zwei Möglichkeiten: (1) durch den Aufbau und Erwerb von Schemata und (2) durch die Automatisierung von Prozessen die ursprünglich unsere ständige kognitive Aufmerksamkeit benötigten. Als Anbieter von Lerninhalten und als Designer von Lernarrangements stellt sich dabei immer wieder die Frage: „Wie sollen Lehrinhalte aufbereitet und dargestellt werden?“ Und in einem weiteren Schritt auch: „Kann durch diese Darstellung gut gelernt werden, wer kann damit gut lernen und wie kognitiv belastend ist Lernen mit diesem Material?“

Gerade die Frage nach der kognitiven Anstrengung beim Lernen ist Thema dieser Arbeit. Enttäuschend hierbei ist immer wieder, das die kognitive Belastung, wie sie in der *Cognitive Load Theory* beschrieben wird bisher nur unzureichend erfasst werden konnte. Im Speziellen ist daher die Entwicklung eines Fragebogens zur differenzierten Erfassung der kognitiven Belastung beim Lernen Ziel der vorliegenden Arbeit. In einem weiteren Schritt soll dann noch geklärt werden, welchen Vorteil eine differenzierte Messung bzw. Erfassung der kognitiven Belastung beim Lernen für das Design von Lernarrangements und die Aufbereitung von Lehrinhalt hat.

Dennoch muss natürlich gesagt werden, dass die kognitive Belastung nicht alleine für Lernerfolg verantwortlich gemacht werden kann. Basierend auf z.B. dem INVO-Modell (Hasselhorn & Gold, 2013) gibt es noch sehr viel mehr Faktoren die Lernen beeinflussen können und im Modell von Hasselhorn und Gold als ineinandergreifende Zahnräder dargestellt sind, darunter (1) Motivation und Selbstkonzept, (2) Volition und lernbegleitende Emotionen, (3) Vorwissen, (4) kognitive Strategien und metakognitive Regulation und (5) die selektive Aufmerksamkeit und das Arbeitsgedächtnis. Die Annahmen der Cognitive Load Theory lassen sich vor allem mit den kognitiven Zahnrädern in Verbindung bringen, aber auch die motivational-volitionalen Zahnräder sind nicht unbeteiligt, aus diesem Grund scheint sie mir besonders relevant für erfolgreiches Lernen zu sein. Dabei kann es sowohl sein, das die kognitive Belastung beim Lernen sich auf Aspekte der Zahnräder auswirkt oder von diesen beeinflusst wird.

Als Kernelement wird in Kapitel 2 die *Cognitive Load Theory* dargestellt und die

dahinterstehenden Gedächtnismodelle werden erklärt. Basierend auf diesen theoretischen Grundlagen werden in einem weiteren Schritt die unterschiedlichen Arten von kognitiver Belastung beschrieben. Zudem wird auf die Kontroverse rund um die Anzahl der Arten kognitiver Belastung eingegangen und diese im Rahmen von Instruktionsdesignfragestellungen betrachtet.

In einem letzten Abschnitt in Kapitel 2 wird noch genauer auf die Messung der kognitiven Belastung beim Lernen eingegangen. Es werden verschiedene Ansätze beschrieben und diskutiert sowie deren Vor- und Nachteile herausgearbeitet.

Abschnitt II dieser Arbeit beinhalte die zu dieser kumulativen Dissertation gehörenden Forschungsartikel und Tagungsbeiträge.

Kapitel 3 enthält den Artikel „Development and Validation of Two Instruments Measuring Intrinsic, Extraneous, and Germane Cognitive Load“ von Klepsch, Schmitz und Seufert (2017) welcher ausführlich die Fragebogenentwicklung beschreibt und die Ergebnisse des Einsatzes desselben in kurzen artifiziellen Lernaufgaben darstellt. Teilergebnisse zu den im Artikel vorgestellten Studien wurden bereits auf verschieden Tagungen vorgestellt, darunter auf der Tagung der European Association for Research on Learning and Instruction (EARLI) 2011 in Exeter (UK) (Pichler & Seufert, 2011) und der International Cognitive Load Theory Conference (ICLTC) 2012 in Tallahassee (USA) (Klepsch & Seufert, 09.-11.04.2012). Zudem wurden Vorträge dazu auf der Arbeitsgruppentagung GöMaEr 2012 in Wuppertal (DE) sowie 2019 in Saarbrücken (DE) gehalten.

Kapitel 4 beschäftigt sich anschließend mit dem praktischen Einsatz des Fragebogens. Der Artikel „Understanding instructional design effects by differentiated measurement of intrinsic, extraneous, and germane cognitive load“ von Klepsch und Seufert (2020) enthält 6 Studien welche der Validierung des Fragebogens dienen und insbesondere die Relevanz der differenzierten Messung der kognitiven Belastung beim Lernen für Instruktionsdesignfragestellungen deutlich machen. Alle sechs Studien wurden entwickelt um Unterschiede in der kognitiven Belastung beim Lernen zu erzeugen. Studie 3 dieser Serie wurde zudem noch erweitert, die Ergebnisse dazu wurden auf der EARLI 2013 in München (DE) vorgetragen (Klepsch, Kempter & Seufert, 2013a). Das selbe gilt für Studie 4 dieses Artikels, auch hier sind weiterführende Analysen auf der 14. Fachgruppentagung Pädagogische Psychologie 2013 in Hildesheim (DE) berichtet worden (Klepsch, Westphal & Seufert, 2013). Passend zu Studie 1 und 2 des Artikels wurde eine ähnliche Studie auf der EARLI 2017 in Tampere (FIN) vorgestellt (Seufert, Klepsch & Westphal, 2017). Im weiteren Verlauf zeigt der Artikel von Rogers et al. (2014) den Einsatz des Fragebogens in einer realistischen Lernumgebung. Bezugnehmend darauf ist zudem der Beitrag von Klepsch, Könings, Weber und Seufert (25.-27.08.2014) auf der Tagung der EARLI Special Interest Group „Comprehension of Text and Graphics“ in Rotterdam (NL) eingefügt, welcher zwei Gruppen der Studie genauer betrachtet und die technische Entwicklung der Lernumgebung außen vor lässt.

Abschließend werden in Abschnitt III dieser Arbeit die Artikel zusammengefasst, sowie der theoretische Mehrwert, die methodischen Stärken und Schwächen und die praktischen Implikationen dargestellt.

2. Cognitive Load Theorie

Die *Cognitive Load Theorie* (CLT; Chandler & Sweller, 1991) ist nun seit fast drei Jahrzehnten nicht mehr aus der Forschung zum Lernen und dem Instruktionsdesign wegzudenken. Anfänglich waren die einzelnen empirischen Befunde von Sweller und Kollegen (Beispielhaft siehe: Mawer & Sweller, 1982; Sweller, Mawer & Howe, 1982; Sweller, Mawer & Ward, 1983; Sweller, 1983; Owen & Sweller, 1985; Sweller & Cooper, 1985; Cooper & Sweller, 1987) noch nicht in ein theoretisches Modell eingebunden und standen lose nebeneinander. Mit Anfang der 90er Jahre änderte sich dies und die CLT entstand. Die CLT fokussiert dabei auf objektiven Charakteristiken der Lernaufgabe bzw. des Lernmaterials und wie diese die kognitive Belastung und damit das Lernen beeinflussen (Moreno & Park, 2010). Einzig das Vorwissen von Lernenden ist als Lernereigenschaft explizit in der CLT genannt und berücksichtigt (Kalyuga, Chandler & Sweller, 1998). Ein großer Kritikpunkt an der Theorie ist daher, das weitere Lernereigenschaften, die den Lernerfolg vorhersagen können (z.B. kognitive Fähigkeiten, Selbstregulation oder Motivation) nicht weiter berücksichtigt werden (siehe auch kognitiv-affektives Modell des Lernens mit Medien; Moreno, 2005). Prinzipiell nimmt die CLT an, dass Lernen immer mit mentaler Belastung verbunden ist, außerdem geht die CLT von einem Arbeitsgedächtnis mit begrenzter Kapazität und einem Langzeitgedächtnis mit quasi unbegrenzter Kapazität aus (Sweller, 2005). Im Langzeitgedächtnis werden sämtliche erlernten Informationen in Form von Schemata gespeichert (Sweller, van Merriënboer & Paas, 1998). Das Langzeitgedächtnis ist demnach unser dauerhaftes Speichersystem im Gehirn, wie und ob Vergessen stattfindet ist noch nicht restlos geklärt, es existieren jedoch einige Vergessenstheorien (z.B. Spurenverfallstheorie, Interferenztheorie, Fehlen geeigneter Abrufreize; vergl. Mietzel, 2017).

Wie bereits erwähnt, nimmt die CLT das Arbeitsgedächtnis mit beschränkter Kapazität an. Basierend auf den Theorien von Baddeley (1986) und Atkinson und Shiffrin (1971) sowie Studien über die Arbeitsgedächtniskapazität (vergleiche Miller, 1956b; Baddeley, 1994) geht die CLT davon aus, das effektives Lernen nur dann möglich ist, wenn die beschränkte Kapazität nicht überschritten wird.

2.1. Lernen und die kognitive Architektur

Lernen selbst hat immer zum Ziel Informationen ins Langzeitgedächtnis zu übertragen. Dabei geschieht Lernen häufig unbewusst und beiläufig (Hasselhorn & Gold, 2017). Ganz allgemein kann Lernen als ein Prozess definiert werden, „der in einer relativ konsistenten Änderung des Verhaltens oder des Verhaltenspotentials resultiert und auf Erfahrung basiert“ (Zimbardo & Gerrig, 2008, Seite 192). Lernprozesse

müssen sich dabei nicht in konkret beobachtbarem Verhalten äußern. Im Kontext von Schule, Hochschule, Weiterbildung, etc. kommt zusätzlich hinzu, dass wir uns wünschen, dass unser Lernen auch erfolgreich ist - z.B. in einer guten Note resultiert. Für Lehrende und Instruktionsdesigner stellt sich in diesem Zusammenhang immer wieder die Frage, wie Lernmaterial auszusehen hat um erfolgreiches Lernen zu unterstützen und wie eine kognitive Überlastung vermieden werden kann.

Doch wie funktioniert nun Wissenserwerb, als eine der wichtigsten Zieldimensionen von Bildungsprozessen (Renkl, 2015)? An dieser Stelle kommt unser Gedächtnis ins Spiel. Unser Gedächtnis nimmt Informationen auf, verarbeitet, transformiert und organisiert sie und legt sie, sofern Lernen gelingt, für den späteren Wiederabruf ab. Um alle diese Prozesse erfolgreich zu bewältigen muss die Struktur unseres Gedächtnisses berücksichtigt werden. Auch die CLT berücksichtigt die im folgenden beschriebenen Theorien und Modelle (Clark, Nguyen & Sweller, 2006).

2.1.1. Drei-Speicher-Modell von Atkinson & Shiffrin

Im Drei-Speicher-Modell von Atkinson und Shiffrin (1971) werden drei Strukturen unterschieden (siehe Abbildung 2.1): (1) die sensorischen Register, (2) das Kurzzeit- bzw. Arbeitsgedächtnis und (3) das Langzeitgedächtnis. Jeder Reiz und jede Information aus der Umwelt (auditiv, visuell, taktil, etc.) wird zuerst über die sensorischen Register aufgenommen und für sehr kurze Zeit als exakte sensorische Kopie festgehalten. Erst wenn einer Information Aufmerksamkeit zuteil wird, gelangt sie ins Arbeitsgedächtnis und kann dort im aktiven Bewusstsein gehalten werden. Über einen begrenzten Zeitraum und mit begrenzter Kapazität können Informationen im Arbeitsgedächtnis wiederholt, verarbeitet, manipuliert, organisiert und ineinander integriert werden. Je nachdem wie lange Informationen im Arbeitsgedächtnis aufrecht erhalten werden können, können diese mehr oder weniger erfolgreich ins Langzeitgedächtnis übertragen werden. Informationen, die nur einmalig und kurz im Arbeitsgedächtnis verarbeitet werden, werden i.d.R. weniger gut mit Informationen aus dem Langzeitgedächtnis verknüpft und werden damit schlechter bis gar nicht erinnert.

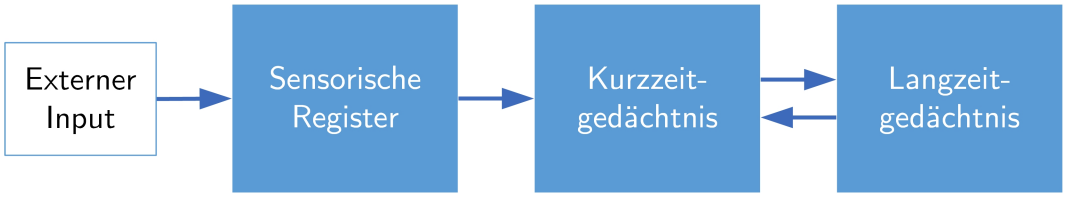

Abbildung 2.1.: Drei-Speicher-Modell nach Atkinson und Shiffrin (1971), übersetzt ins Deutsche

Die CLT fokussiert dabei rein auf das Arbeitsgedächtnis und das Langzeitgedächtnis. Die sensorischen Register sind in der Theorie außen vor. Dabei geht es in der CLT vor allem darum, die kognitive Belastung im Arbeitsgedächtnis zu kontrollieren

bzw. zu reduzieren um Überlastung zu vermeiden (Clark et al., 2006). Erst wenn Lernen im Arbeitsgedächtnis erfolgreich ist, kann das Erlernte in Form von Schemata ins Langzeitgedächtnis übertragen werden (NSW Department of Education, 2017). Zudem können diese Schemata, welche nun das Vorwissen enthalten, zurück ins Arbeitsgedächtnis geholt werden. Bei der Verarbeitung neuer Informationen kann dann an dieses Vorwissen angeknüpft werden und die Schemata können so erweitert werden.

Ein der CLT ähnliches Konzept ist die Perceptual Load Theorie von Lavie (1995): In dieser Theorie kommen die sensorischen Register noch ins Spiel. Sind die attentionalen Anforderungen an eine zu bearbeitende Aufgabe gering so landen auch irrelevante Distraktoren im Sensorischen Register, da entsprechende Kapazitäten frei sind. In diesem Fall wird von geringem percetual load ausgegangen. Die Theorie nimmt weiter an, das diese irrelevanten Distraktoren nun natürlich auch weiterverarbeitet werden können und so eine Antwortinterferenz mit der eigentlichen Aufgabe verursachen können. Beansprucht eine Aufgabe jedoch die vollständige Aufmerksamkeit, können Distraktoren nicht mehr im sensorischen Register verarbeitet werden und haben so auch keine Chance mit Aufmerksamkeit belegt zu werden (Müller & Krummenacher, 2012). In den klassischen Experimenten zur Perceptual Load Theorie, ist die Aufgabe meist eine Targetziel zu identifizieren und entsprechend zu reagieren. Lernen funktioniert vielfach jedoch nicht so. Konzentrieren sich Lernende auf ihre Aufgabe und fordert diese aufgrund ihrer Komplexität ausreichend Aufmerksamkeit sind Distraktoren wie von der Perceptual Load Theorie angenommen irrelevant, da sie gar keine Chance haben, mit Aufmerksamkeit belegt zu werden. Ist die Aufgabe für einen Lernenden jedoch sehr einfach, kommt die Perceptual Load Theorie zum Tragen, und auch gänzlich Aufgaben-irrelevante Informationen können mit Aufmerksamkeit belegt werden (z.B. eine Socke, die aus dem Augenwinkel auf dem Boden entdeckt wird, kann dazu führen, dass im Arbeitsgedächtnis die nächste Wäscheladung schon mit geplant wird). Diese sind jedoch für den Lernprozess nach der CLT irrelevant, dieses *Mind-Wandering* (Smallwood & Schooler, 2014) würde den Lernprozess entweder ins stocken geraten lassen oder ihn ganz abbrechen lassen. In diesem Fall wäre die kognitive Belastung beim Lernen nach der CLT nicht mehr relevant, da der Lernprozess beendet wurde.

2.1.2. Duale Kodierung nach Paivio

Zur selben Zeit postulierte Paivio (1971, 1986) mit der Theorie der dualen Kodierung, dass verschiedene Repräsentationen auch verschieden verarbeitet und gespeichert werden. Bildliche Informationen (Imagene) und sprachliche Informationen (Logogene) werden in unabhängigen, aber auf vielfältige Weise miteinander verbundenen Systemen verarbeitet und gespeichert. Diese Annahme inkludiert, dass die Verarbeitungssysteme sich funktional unterscheiden, da sie mit Inhalten unterschiedlicher Art umgehen müssen (siehe Abbildung 2.2): (1) Verbale Informationen werden sequenziell und nach logisch analytischen Regeln verarbeitet und in Form von Wortmarken (Logogene) abgespeichert. (2) Piktorale Informationen werden holistisch-

analog verarbeitet und in Form von mentalen Bildern (Imagene) abgespeichert.

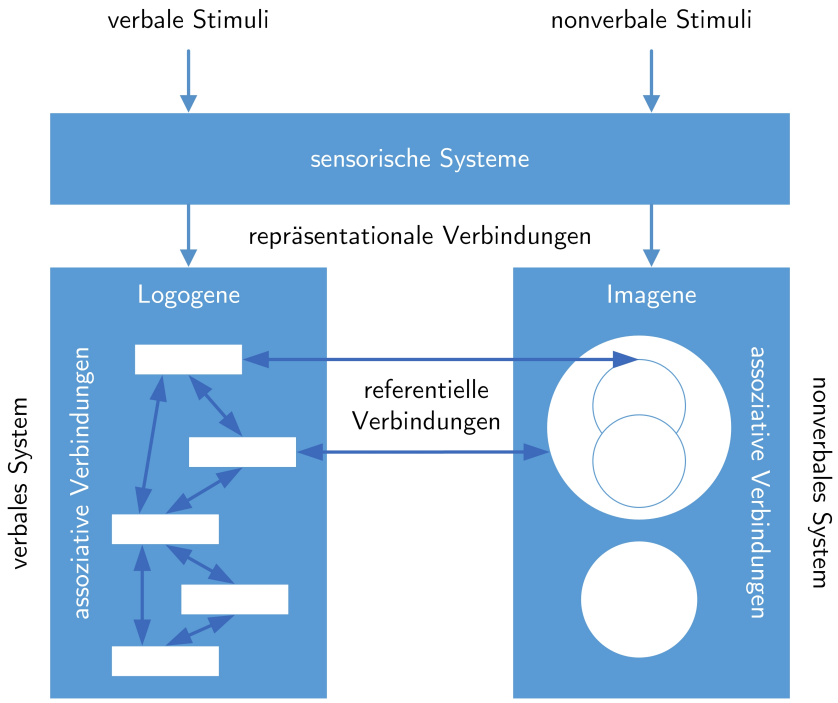

Abbildung 2.2.: Duale Kodierung nach Paivio (1986), übersetzt ins Deutsche

Die beiden Systeme sind in vielfältiger Weise miteinander verbunden und können sich gegenseitig aktivieren. Paivio (1986) beschreibt zudem den *Picture-Superiority-Effect*, welcher darauf beruht, dass Bilder einen höheren Wiedererkennungseffekt haben und damit einfacher erlernt bzw. gemerkt und abgerufen werden können. Dies liegt vor allem daran, dass Bilder, in Einklang mit der Theorie der dualen Kodierung, leichter und schneller doppelt kodiert werden können. Was man sieht, kann man in der Regel auch in Worte fassen.

Für Instruktionsdesignfragestellungen und das Erstellen von Lernmaterialien bedeutet dies im Speziellen, dass das Darbieten von bildhaften Repräsentationen das Lernen unterstützen müsste und sich dies auch in einer reduzierten kognitiven Belastung zeigen sollte.

2.1.3. Arbeitsgedächtnismodell von Baddeley

Die von Paivio (1971) getroffene Annahme unterschiedlicher Verarbeitungssysteme geht auch in das Arbeitsgedächtnismodell von Baddeley (1986) ein. In diesem geht Baddeley (1986) davon aus, dass das Arbeitsgedächtnis in mehrere Subsysteme aufgeteilt ist (siehe Abbildung 2.3). Die zentrale Exekutive sorgt für die Koordination der phonologischen Schleife, des räumlich-visuellen Notizblocks und des episodischen Puffers. Zudem ist sie für die Integration der Informationen aus dem

Langzeitgedächtnis verantwortlich. Dem Modell nach verarbeitet die phonologische Schleife alle sprachlichen Informationen, der räumlich-visuelle Notizblock alle visuellen Informationen und der episodische Puffer wird multimodal angenommen und speichert Informationen kurzzeitig in Form von Episoden. Da Lerninhalte i.d.R. sprachlich oder visuell dargeboten werden sind die phonologische Schleife und der räumlich-visuelle Notizblock fürs Lernen am relevantesten.

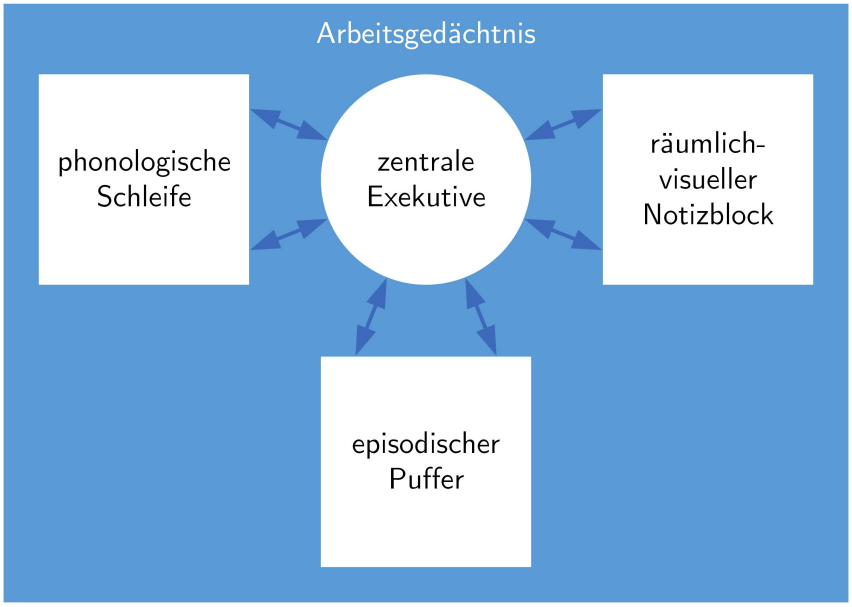

Abbildung 2.3.: Arbeitsgedächtnismodell nach Baddeley (1986), übersetzt ins Deutsche

Durch das im Zusammenhang mit der CLT immer wieder genannte *Modalitätsprinzip* (Lernen mit Bild und zugehörigem auditiven Text ist erfolgreicher als Lernen mit Bild und zugehörigem visuellen Text; Ginns, 2005; Sweller, van Merriënboer & Paas, 2019), zeigt sich, dass die unterschiedlichen Verarbeitungssysteme auch in die CLT-Theorie einfließen, bzw. von ihr beachtet werden.

Neben dem Komponentenmodell, wie Baddeley's Arbeitsgedächtnismodell auch genannt werden kann, gibt es noch weitere Arbeitsgedächtnismodelle (Miyake & Shah, 2007). Exemplarisch wird hier noch das Prozessmodell von Cowan (Cowan, 1995) genannt. Es stützt sich auf das neuronale Netzwerk und geht davon aus, dass das Arbeitsgedächtnis jene Anteile des Langzeitgedächtnisses darstellt, die sich gerade (und vorübergehend) in einem aktivierten Zustand befinden. Teile der aktivierten Informationen sind dann im Fokus der Aufmerksamkeit. Das Prozessmodell von Cowan nimmt an, dass die Anzahl aktivierter Informationen kapazitätsbeschränkt ist und die gerade im Fokus der Aufmerksamkeit stehenden Informationen einem Zeitlimit unterliegen (Cowan, 2007). Somit würde die CLT auch in diesem Modell greifen und ließe sich entsprechend übertragen.

Wie bereits kurz erwähnt unterliegt unser Arbeitsgedächtnis einer Kapazitäts-

beschränkung. Abschließend muss daher noch die Frage gestellt werden, wie viel Kapazität wir denn fürs Lernen im Arbeitsgedächtnis zur Verfügung stellen können. Da die Arbeitsgedächtniskapazität individuell unterschiedlich ausfällt (Cowan, 2010) muss schließlich auch davon ausgegangen werden, dass eine hohe Arbeitsgedächtniskapazität deutlich Lernförderlicher sein müsste als eine geringe Arbeitsgedächtniskapazität.

2.1.4. Arbeitsgedächtniskapazität

In den Annahmen der CLT zum Arbeitsgedächtnis wird davon ausgegangen, dass die Kapazität des Arbeitsgedächtnisses beschränkt ist. Oft wird die Arbeitsgedächtniskapazität mit 7±2 *Chunks* (Miller, 1956a) beziffert. Diese „magische“ Zahl hält sich hartnäckig obwohl Miller bereits seine Leser ermahnte sein rhetorisches Mittel nicht für bare Münze zu nehmen (Miller, 1956a) und dies später noch einmal bekräftigte (Miller, 1989). Prinzipiell gilt jedoch, dass wir nur eine begrenzte Menge von Informationen aktiv im Arbeitsgedächtnis halten können. Können wir einzelne Informationen verknüpfen, und so sinnvolle zusammengehörige Einheiten erschaffen, entstehen Chunks durch die wir „virtuell“ diese Kapazitätsbeschränkung unseres Arbeitsgedächtnisses erweitern. Offenbar ist Lernen im allgemeinen ein großes Trainingscamp für *Chunking*, da jede neue Information Chunking prinzipiell unterstützen kann. Auch in Millers Studien war Chunking möglich, was darauf schließen lässt, dass die Kapazitätsgrenze eher überschätzt wurde. Neuere Studien (Cowan, 2001, 2010; Saults & Cowan, 2007) schätzen die „magische“ Zahl daher auf 4±1 Chunks. Dies hat natürlich Auswirkungen auf das Lernen. Wir können annehmen, dass beispielsweise Verstehen - und damit die Bildung eines korrekten mentalen Modells - erst dann stattfinden kann, wenn alle für den Sachverhalt wichtigen Informationen und Elemente simultan im Arbeitsgedächtnis verarbeitet werden können.

Die CLT folgt dieser Annahme der begrenzen Kapazität sehr direkt und unterteilt die kognitive Belastung beim Lernen, die im Arbeitsgedächtnis wirkt, auch noch auf unterschiedliche Quellen auf.

2.2. Arten kognitiver Belastung

Die im Arbeitsgedächtnis zu verarbeitenden Informationen setzen sich in der traditionellen Sicht der CLT aus drei Quellen kognitiver Belastung zusammen: dem intrinsic cognitive Load, dem extraneous cognitive Load und dem germane cognitive Load (Moreno & Park, 2010). Eine überarbeitete Version der Theorie sieht nur noch zwei Quellen kognitiver Belastung vor, dazu jedoch später mehr (siehe Abschnitt 2.3.

2.2.1. intrinsic cognitive Load (ICL)

Die intrinsische Belastung entsteht durch das Lernmaterial selbst, sie resultiert aus der Schwierigkeit und der Komplexität des zu bearbeitenden Lernmaterials (Ayres,

2006). Die Elementinteraktivität kann dabei als „Maß“ der Schwierigkeit und Komplexität angesehen werden (Sweller & Chandler, 1994; Paas, Renkl & Sweller, 2003). So ist die Elementinteraktivität beim Lernen einzelner Vokabeln in einer Fremdsprache deutlich geringer als die Elementinteraktivität beim Erlernen der dazugehörigen Grammatik. Während die einzelnen Vokabeln isoliert voneinander erlernt werden können, ist Grammatik nur durch deutlich mehr Elemente, die verarbeitet werden müssen, zu erlernen. Als Faustregel kann demnach davon ausgegangen werden: Je höher die Elementinteraktivität, desto höher auch der ICL. Beim ICL wird auch die einzige in der Theorie berücksichtigte Lernereigenschaft wichtig: Das Vorwissen und die damit verbundene Fähigkeit zum Chunking. Damit wird es im einzelnen mitunter schwer, die Elementinteraktivität für den einzelnen Lerner überhaupt festzumachen: Während Experten auf einem Gebiet einen komplexen Sachverhalt als Chunk im Arbeitsgedächtnis verarbeiten können, ist dies für Novizen nicht möglich und sie müssen jedes Element einzeln verarbeiten (Sweller et al., 1998). Da eine Verringerung des ICL oft nur möglich ist indem man Abstriche in der Komplexität des Lernmaterials macht, was bei einem gegebenen Lern- bzw. Lehrziel oft schwer ist, ist es für Instruktionsdesignfragestellungen wichtig das Vorwissen der Lernenden beim Erstellen der Lernmaterialien zu berücksichtigen. Dabei geht es weniger um die Menge des zu lernenden Inhaltes, diese muss schließlich dem Lehrziel entsprechen, als mehr um den Aufbau des zu erlernenden Inhaltes. So werden beispielsweise beim *Pre-training Principle* (Mayer, 2009; Renkl & Scheiter, 2017) wichtige Begriffe bereits vor der Bearbeitung des eigentlichen Lerninhaltes präsentiert, sodass die Elementinteraktivität und damit der ICL für Lernende, die mit diesen Begriffen noch nicht vertraut sind, gesenkt wird.

2.2.2. extraneous cognitive Load (ECL)

Die Darstellung und Gestaltung des Lernmaterials bedingt den ECL. Wenn Lernmaterial viele irrelevante Informationen beinhaltet, führt dies zu einer unnötig höheren Belastung, da der Lernende alle unbenötigten Informationen ausblenden muss. Außerdem führen schlechtes Design und ungünstige Gestaltung des Lernmaterials zu höherem ECL, da die schlechte Darstellung ausgeglichen werden muss. Die CLT propagiert, dass um ECL zu vermeiden der Lernstoff in einer möglichst optimalen Darstellung und Aufbereitung vorliegen sollte. Aus diesem Grund sind gerade in Bezug auf die extrinsische Belastung viele Designprinzipien entstanden, die beschreiben wie Lernmaterial dargestellt werden sollte. Darunter fallen zum Beispiel das *Multimediaprinzip* (Mayer, 2009; Butcher, 2014), welches besagt, dass Lernen mit Text und relevantem Bild besser gelingt als ohne das Bild. Auch das *Split-Attention-Prinzip* (Sweller & Chandler, 1994; Florax & Ploetzner, 2010; Schroeder & Cenkci, 2018), fällt in diese Kategorie: Werden Text und Bild zusätzlich sinnvoll integriert und nicht separiert dargestellt, kann der Lernerfolg gesteigert werden. Denn durch die Integration der beiden Repräsentationen können Suchprozesse verhindert werden, die ansonsten ECL erzeugen würden. Durch das *Seductive Details Prinzip* (Rey, 2012; Sundararajan & Adesope, 2020) wird auch noch klar, dass nur relevante Bilder

den Lernerfolg steigern. Haben Bilder nur eine rein dekorative Funktion sind sie eher lernhinderlich.

2.2.3. germane cognitive Load (GCL)

Ein weiterer Teil der kognitiven Belastung ist der GCL. Lange wurde davon ausgegangen, dass die kognitive Belastung möglichst gering gehalten werden sollte, mit Einführung von GCL wurde die Theorie jedoch um eine Belastungsart ergänzt, die positiv für den Lernenden ist (Moreno & Park, 2010). Der GCL stellt jene Aktivitäten dar, die dabei helfen das Lernmaterial zu verstehen und effektive Schemata im Langzeitgedächtnis zu bilden (Sweller et al., 1998; Bannert, 2002). Eine vertiefte Auseinandersetzung mit dem Lerninhalt kann also eine lernförderliche Aktivität sein.

Angenommen, wenn auch empirisch noch nicht ausreichend bewiesen, wird, dass der GCL nur dann Kapazitäten beanspruchen kann, wenn die vorhandene Kapazität nicht bereits durch ICL und ECL beansprucht wird (siehe auch Additivitätshypothese; Zander, 2010; Park, 2010)). GCL kann insbesondere von Seiten der Lehrenden und Instruktionsdesigner durch Hilfestellungen wie *Prompts* (vergleiche auch Bannert, 2009; Wirth, 2009; Bisra, Liu, Nesbit, Salimi & Winne, 2018) gefördert werden: z.B. wird versucht mit Hilfe aktivierender Lernfragen o. ä. gezielt den GCL zu erhöhen.

In vielen Fällen kann der GCL jedoch nicht direkt manipuliert werden, da der eingesetzte Aufwand den Lernende betreiben immer noch in der Verantwortung der Lernenden selbst liegt und Lehrende und Instruktionsdesigner Lernende schwer zwingen können sich anzustrengen und kognitive Ressourcen einzusetzen. Eine Ausnahme könnte hier der *Disfluency Effekt* bilden (Diemand-Yauman, Oppenheimer & Vaughan, 2011), bei diesem wird die investierte Anstrengung erhöht, da das Material durch erschwerte Lesbarkeit als subjektiv schwerer wahrgenommen wird.

2.3. Kritik an der CLT oder wie viele Arten kognitiver Belastung gibt es nun wirklich?

Im Laufe der letzten Jahre wurde vielfach diskutiert ob die Unterteilung in intrinsic, extraneous und germane cognitive Load richtig und empirisch haltbar ist. Gerade im Buch „Cognitive Load Theory“ von Sweller, Ayres und Kalyuga (2011) selbst, wird der GCL beispielsweise außen vor gelassen. Einer der größten Kritikpunkte ist tatsächlich die Unterscheidung der verschiedenen Formen kognitiver Belastung. Während ICL und ECL möglichst reduziert werden sollen und hoher ICL bzw. hoher ECL größtenteils als lernhinderlich eingestuft wird, verhält es sich mit dem GCL genau umgekehrt, hoher GCL wird als lernförderlich betrachtet (Schnotz & Kürschner, 2007; de Jong, 2010). Basierend auf diesem konzeptionellen Unterschied zwischen den unterschiedlichen Belastungsarten regten Kalyuga (2011), Ayres (2011) und Sweller et al. (2011) an, das Konzept des GCL zu überdenken. Sie alle argumentieren damit,

dass GCL eher als eine Art Ressource im Arbeitsgedächtnis aufgefasst werden könne um den intrinsischen Anforderungen des Materials gerecht zu werden. Damit wird aus überdenken eher ein umbenennen, denn es bleibt offen ob nicht deutlich mehr Ressourcen zur Verfügung gestellt werden können, als alleine für die Verarbeitung des Inhaltes nötig sind. Damit ergibt sich insgesamt folgendes theoretisches Konstrukt (Sweller et al., 2011): Die Komplexität des Lerninhaltes stellt den ICL dar, die gute oder schlechte Darstellung des Materials spiegelt sich im ECL wieder. Beide Arten der kognitiven Belastung müssen im Arbeitsgedächtnis verarbeitet werden. Um dies zu ermöglichen müssen im Arbeitsgedächtnis Kapazitäten zur Verarbeitung bereitgestellt werden. Die für die Verarbeitung des unerwünschten ECL bereitgestellten Kapazitäten können als *extraneous resources* (ER) bezeichnet werden. Die für die inhaltliche Verarbeitung bereitgestellten Kapazitäten können als *germane resources* (GR) bezeichnet werden. Dieser Sachverhalt ist in Abbildung 2.4 dargestellt, dabei wurde jedoch berücksichtigt, dass die theoretischen Annahmen nicht ausschließen, dass mehr GR im Arbeitsgedächtnis zur Verfügung gestellt werden können, als für die reine Inhaltsverarbeitung nötig wären. Diese zusätzlichen Arbeitsgedächtniskapazitäten könnten Schemakonstruktion, Chunkbildung und den Versuch, den Inhalt tiefer zu verstehen, fördern. Damit ergäben sich für Instruktionsdesignfragestellungen wieder drei zu messende Konstrukte: (1) ICL, (2) ECL, (3) GR - wobei GR nicht der Definition von GCL widerspricht.

Abschließend bleibt an dieser Stelle daher nur zu sagen, dass für Instruktionsdesignfragestellungen, die Bezeichnung der einzelnen Konstrukte irrelevant ist, wichtig ist die Tatsache, dass immer noch kontrolliert werden muss, welche kognitive Belastung auf den Lerninhalt selbst, welche auf das Design des Materials bzw. der Lernumgebung und welche auf lernförderliche Maßnahmen durch Lehrende oder Lernende selbst zurückzuführen ist.

Doch wie kann die kognitive Belastung beim Lernen nun überhaupt erfasst und gemessen werden?

2.4. Methoden um kognitive Belastung zu messen

Das Messen der einem Lernprozess zugehörigen kognitiven Belastung stellt Forscher immer wieder vor eine Herausforderung. Die Kategorisierung verschiedener Messmethoden wird zudem unterschiedlich gehandhabt: Unterschieden werden subjektive, aufgaben- bzw. leistungsbasierte und physiologische Indikatoren (Wierwille & Eggemeier, 1993), subjektive und objektive Indikatoren, die zusätzlich in indirekt und direkt unterteilt werden (Brünken, Plass & Leutner, 2003), oder auch Beurteilungsskalen, psychophysiologische Messungen und Dual-Task-Techniken (Paas, Tuovinen, Tabbers & van Gerven, 2003). Die größte Übereinstimmung ergibt sich in der einfachen Unterscheidung zwischen subjektiven und objektiven Techniken (vergleiche z.B. Brünken, Seufert & Paas, 2010) wobei eine weitere Unterteilung nicht mehr auf beide Kategorien gleich angewandt wird.

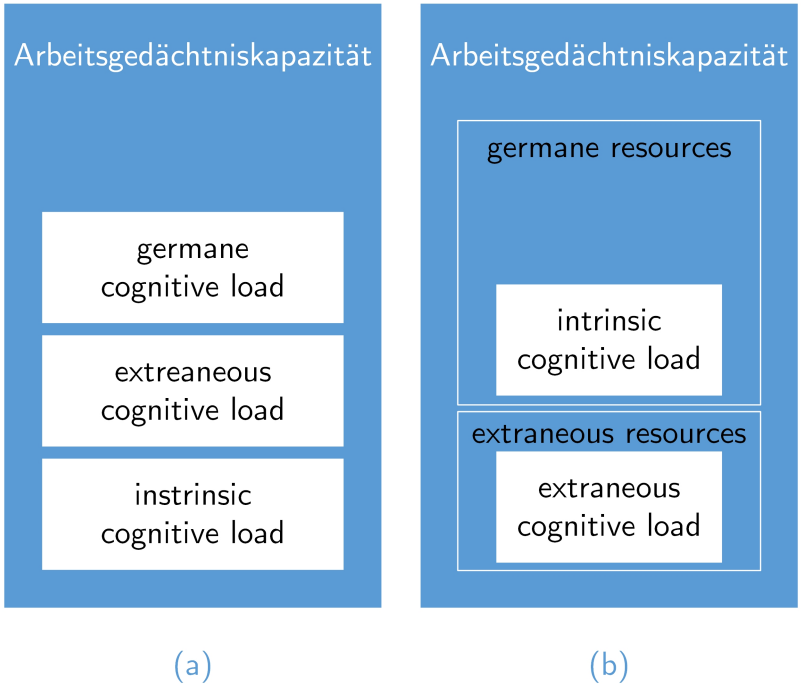

Abbildung 2.4.: (a) Traditionelle Sicht der CLT in Anlehnung an Moreno und Park (2010), ICL, ECL und GCL werden als Quellen für Arbeitsgedächtnisbelastung gesehen (b) Überarbeitete Sicht der CLT in Anlehnung an Sweller et al. (2011), ICL und ECL als Belastungsarten die auf das Arbeitsgedächtnis wirken und dort durch die Bereitstellung von Kapazitäten verarbeitet werden können.

2.4.1. Subjektive Indikatoren

Subjektive Indikatoren sind dadurch gekennzeichnet, dass Lernende selbst ihre Einschätzung zu ihrer kognitiven Belastung beim Lernen abgeben. In der Regel werden retrospektive Fragen, welche vom Lernenden während oder nach der Aufgabenbearbeitung beantwortet werden, hierzu eingesetzt. Obwohl immer wieder angezweifelt wird, dass Personen zu einer adäquaten Introspektion fähig sind (vergl. Schnotz & Kürschner, 2007), gibt es in der Literatur auch empirische Belege, die bestätigen, dass Personen dazu in der Lage sind (Gopher & Braune, 1984).

Die wohl bekannteste und am meisten angewandte Messmethode um die kognitive Belastung beim Lernen oder der Aufgabenbearbeitung zu messen ist die Frage nach der mentalen Anstrengung. Paas (1992) entwickelte dazu ein Item, dieses wurde seither in unterschiedlichen Varianten in der Cognitive Load Forschung eingesetzt. Der genaue Wortlaut wurde im Paper von 1992 (Paas, 1992) zwar nicht veröffentlicht, oft genutzte Item-Texte im Englischen lauten jedoch „I invested ... mental effort“ oder „My invested mental effort was ...“. Im Deutschen wird meist der Wortlaut „Meine mentale Anstrengung war ...“ genutzt. Die Frage wird anschließend auf einer meist 7 bis 9-stufigen Likert Skala mit den Endpunkten „sehr gering“ (very low“) bis „sehr hoch“ („very high“) beantwortet (für einen Überblick siehe Paas, Tuovinen et al., 2003). Obwohl Studien (Paas, van Merriënboer & Adam, 1994; Kester, Kirschner & van Merriënboer, 2004; van Gog & Paas, 2008) zeigen konnten, dass diese unidimensionale Skala reliabel und sensitiv ist und die Anwendung dieser auch weit verbreitet ist (Paas, Tuovinen et al., 2003), werden doch immer wieder kritische Stimmen laut (z.B. de Jong, 2010; Kirschner, Ayres & Chandler, 2011).

Einerseits ist unklar, ob die Messung von kognitiver Belastung beim Lernen durch die Frage nach der mentalen Belastung nun direkt oder indirekt ist: Während Sweller et al. (2011) das Item als direkte Messung von kognitiver Belastung bezeichnen (siehe auch Paas, Renkl & Sweller, 2003; van Gog & Paas, 2008), wird es von Brünken et al. (2003) als indirekt eingeordnet. Zweitere argumentieren damit, dass unklar ist in welcher Relation die angegebene mentale Belastung zur aktuellen kognitiven Belastung steht.

Andererseits kann ein Item einfach nicht drei Arten von kognitiver Belastung erfassen, dazu kommt im Zusammenhang mit z.B. Lernerfolgsmessungen der Rückschluss auf die eventuell ausschlaggebende Art von kognitiver Belastung - es entsteht ein Ringschluss (Kirschner et al., 2011). Ein Beispiel: Zwei Studierende berichten ein ähnlich hohes Level mentaler Belastung, dies für sich alleine ist ein Ergebnis, interessant wird es aber erst, wenn der Lernerfolg mit betrachtet wird: Einer hat einen hohen Lernerfolg - seine Anstrengung scheint erfolgreich gewesen zu sein, es könnte von hohem GCL ausgegangen werden. Der andere hat einen niedrigen Lernerfolg, für ihn könnte nun der Inhalt zu komplex gewesen sein (hoher ICL) oder die Darstellung ungünstig gewesen sein (hoher ECL).

Ähnliche Items, wie beispielsweise die Frage nach der empfundenen Schwierigkeit (DeLeeuw & Mayer, 2008), sind den selben Kritikpunkten unterworfen.

Zudem haben bereits Forscher versucht, die drei Arten kognitiver Belastung beim

Lernen differenziert zu messen, dies ist mehr oder weniger gut gelungen, wurde oft aber nur für eine Studie ausprobiert oder gar auf ein spezielles Lernsetting zugeschnitten. Da die Entwicklung eines differenzierten Fragebogens wichtiger Bestandteil dieser Arbeit ist, wird im Abschnitt 2.5 speziell auf die Möglichkeit, die drei Arten kognitiver Belastung differenziert zu messen, eingegangen.

2.4.2. Objektive Indikatoren

Objektive Indikatoren können Aufgabencharakteristika und Leistungsindikatoren, physiologische Indikatoren oder verhaltensbasierte Indikatoren sein.

Aufgaben- bzw. leistungsbasierte Indikatoren

Unter diese Messmethoden fallen alle Techniken die objektiv Aufgabencharakteristika und Leistungsindikatoren erfassen. Wenn davon ausgegangen wird, dass Lernen erschwert ist wenn das Arbeitsgedächtnis überlastet ist - also eine zu hohe kognitive Belastung vorliegt - würde sich dies in einem Leistungsabfall bemerkbar machen (Chen et al., 2016). Dabei kann zwischen Indikatoren unterschieden werden, die auf die primäre Lernaufgabe zurückzuführen und solchen, die auf eine sekundäre Aufgabe zurückzuführen sind (Brünken, Steinbacher, Plass & Leutner, 2002). Während z.B. der Lernerfolg oder auch die Lernzeit Indikatoren sind, die auf die primäre Lernaufgabe zurückzuführen sind und die kognitive Belastung nur indirekt erfassen, ist die Leistung in einer Zweitaufgabe (*Dual-Task-Paradigma*; z.B Brünken et al., 2002) logischerweise auf die sekundäre Aufgabe zurückzuführen und misst die kognitive Belastung direkt (Brünken et al., 2002, 2003). Zweitaufgaben können das reagieren auf entsprechende Stimuli sein oder auch das tappen eines Rhythmus. Beide Methoden haben ihre Stärken und Schwächen. So können die auf die primäre Aufgabe zurückzuführenden Indikatoren zwar einfach erfasst werden, da sie sich einfach in Studien integrieren lassen, für sich alleine sind sie jedoch meist nicht sehr aussagekräftig und können oft nur in Kombination mit anderen Messarten sinnvoll die kognitive Belastung beim Lernen erklären (Kirschner et al., 2011). Das Dual-Task-Paradigma wurde bereits in vielen Studien genutzt um die kognitive Belastung beim Lernen zu erfassen (vergl. Park & Brünken, 2015; Brünken, Plass & Leutner, 2004; Brünken et al., 2002), allerdings kann auch diese Methode nur die gesamte kognitive Belastung beim Lernen erfassen und ermöglicht keine Differenzierung der einzelnen Arten kognitiver Belastung.

Physiologische Indikatoren

Physiologische Indikatoren wurden vor allem in den letzten Jahren immer öfter zur Messung der kognitiven Belastung eingesetzt. Physiologische Indikatoren zeichnen sich dadurch aus, dass ihre Veränderungen nicht willentlich ausgelöst werden können, sondern die Folge einer Reaktion auf Umwelteinflüsse sind. Beispiele für physiologische Indikatoren sind Herzfrequenz und -variabilität, Gehirnaktivität, galvanische Hautreaktionen, Hautleitfähigkeit, Blinzelrate und Pupillenveränderungen (ein

Überblick bietet Chen et al., 2016). Durch diese Verfahren kann die kognitive Belastung über längere Zeitintervalle kontinuierlich erfasst werden. Allerdings können *Workload*-Spitzen nicht präzise abgebildet werden und die Verfahren sind im Feld nicht praktisch anwendbar, der Aufwand zur Messung der kognitiven Belastung ist im Vergleich zu anderen Messmethoden deutlich erhöht (Rey, 2009) und auch die Auswertung der Daten deutlich komplexer. Außerdem können auch nicht kognitive Erregungszustände, z.B. Angst, Freude, etc., solche physiologischen Muster erzeugen.

Verhaltensbasierte Indikatoren

Verhaltensbasierte Indikatoren zeichnen sich dadurch aus, dass versucht wird die Menge der kognitiven Belastung anhand von bestimmtem Verhalten abzuschätzen. Hierunter fällt das *Eye-Tracking*. Augenbewegungen können genutzt werden, um die kognitive Belastung beim Lernen zu erfassen, sogenannte Sakkaden und Fixationen geben Aufschluss darüber, welche Lernmaterialien gerade verarbeitet und/oder verknüpft werden. Die Dauer der Sakkaden und Fixationen kann die Höhe der kognitiven Belastung abbilden (Zagermann, Pfeil & Reiterer, 2016). Die Messung verhaltensbasierter Indikatoren, also was macht eine Person, kann teils, je nach gegebenen technischen Möglichkeiten, ebenfalls nur mit größerem Aufwand realisiert werden und auch die Auswertung der Daten ist komplex.

2.4.3. Bewertung der Messmethoden

Je nach Lernumgebung und Intention kann die Art der Erfassung der kognitiven Belastung unterschiedlich günstig oder ungünstig für den Lernprozess und die Aussagekraft sein. Wierwille und Eggemeier (1993) geben folgende zwei Punkte an, welche bei der Auswahl eines passenden Erhebungsinstruments berücksichtigt werden sollten: (1) Sensitivität und (2) Störung des Lernprozesses. Während Sensitivität beschreibt inwieweit ein Messinstrument Unterschiede in der kognitiven Belastung erfassen kann, geht es bei der Störung des Lernprozesses darum, inwieweit die Erfassung der kognitiven Belastung den Lernprozess einer Versuchsperson stört oder unterbricht. Das Dual-Task-Paradigma stellt bezüglich der Störung des Lernprozesses eine Ausnahme da, es soll ja um Ressourcen im Arbeitsgedächtnis mit dem eigentlichen Lerninhalt konkurrieren (Korbach, Brünken & Park, 2017).

Im Folgenden wird, basierend auf den Kriterien von Wierwille und Eggemeier (1993) versucht, die zuvorgenannten Messmethoden einzuordnen.

Subjektive Indikatoren scheinen sensitiv genug zu sein um Unterschiede in der kognitiven Belastung beim Lernen erfassen zu können, je nach Einsatz sind sie mehr oder weniger störend. Wird am Ende der Lerneinheit abgefragt, so wird der Lernprozess nicht gestört. Werden während des Lernprozesses immer wieder eine oder mehrere entsprechende Fragen eingeschoben, erhöht sich logischerweise die Störung des Lernprozesses. Allerdings konnten van Gog, Kirschner, Kester und Paas (2012) zeigen, dass genau dies jedoch einen Unterschied macht und sich somit auf die Sen-

sitivität auswirkt: die subjektive Einschätzung der kognitiven Belastung ist höher, wenn nur einmalig am Ende einer Serie von Aufgaben gefragt wird, als wenn das Mittel der subjektiven Einschätzung der kognitiven Belastung über die einzelnen Aufgabe hinweg berechnet wird.

Aufgaben- bzw. leistungsbasierte Indikatoren stören den Lernprozess gar nicht, da sie i.d.R. im Hintergrund mit erfasst werden oder im Nachhinein berechnet werden können ohne in den Lernprozess selbst einzugreifen. Wie bereits erwähnt stört eine Zweitaufgabe den Lernprozess erheblich, dies ist jedoch gewünscht. Die Zweitaufgabe ist zudem deutlich sensitiver, denn sie erfasst die kognitive Belastung direkter (Brünken et al., 2002). Aufgaben- und leistungsbasierte Indikatoren aus der Primäraufgabe hingegen sind zwar sensitiv, aber ggf. nicht in Bezug auf die kognitive Belastung, so könnte zum Beispiel eine schlechte Leistung in einem Lernerfolgstest auch auf andere Einflussvariablen zurück geführt werden.

Physiologische Indikatoren und verhaltensbasierte Indikatoren stören durch die nötigen Geräte den Lernprozess zum Teil erheblich, da teilweise entsprechende Einschränkungen im Bewegungsradius der Versuchspersonen vorhanden sind. Diese Methoden eignen sich daher meist auch nicht für Feldstudien. Die Methoden sind zwar sensitiv, wie allerdings bereits erwähnt können je nach Methode Spitzen in der kognitiven Belastung teils nicht oder nur verzögert erfasst werden (Rey, 2009).

Basierend auf dieser Einschätzung wären subjektive Indikatoren, den objektiven teils überlegen, dennoch unterliegen alle bisher vorgestellten Methoden einer Einschränkung: Sie alle Messen die kognitive Belastung lediglich als Gesamtkonstrukt, eine differenzierte Betrachtung der einzelnen Arten kognitiver Belastung bleibt aus.

2.5. Differenzierte Messung der kognitiven Belastung

Die vorliegende Arbeit beschäftigt sich mit der Entwicklung eines Fragebogens zur differenzierten Messung der kognitiven Belastung beim Lernen. Differenziert deshalb, da die oben beschriebenen Verfahren lediglich dazu in der Lage sind die gesamte kognitive Belastung zu messen, nicht jedoch die einzelnen Arten kognitiver Belastung: intrinsic cognitive Load, extreanous cognitive Load und germane cognitive Load.

Eine Theorie kann jedoch nur vollständig erforscht und überprüft werden, wenn all ihre Bestandteile und Annahmen erfasst und gemessen werden können. Dies ist durch die vorgestellten objektiven Methoden und auch die bereits vorgestellten subjektiven Methoden nicht vollständig möglich. Werden ICL, ECL und GCL nur als Konglomerat erfasst, kann die Theorie nicht befriedigend weiterentwickelt werden. Die Annahmen wie sich die Arten der kognitiven Belastung auf den Lernerfolg auswirken oder durch Lernmaterialveränderungen verändert werden können nicht überprüft werden.

Mehrere Forscher haben sich bereits an einer differenzierten Messung versucht um ihre Instruktionsdesignfragestellungen bestmöglich zu beantworten und die Forschung an Multimediadesignprinzipien voranzubringen. Im Folgenden werden einige relevante Ansätze kurz genannt. In Klepsch et al. (2017), einem zu dieser Dissertati-

on gehörenden Beitrag (siehe Kapitel 3), wird die Thematik ausführlicher behandelt.

Die ersten Ansätze der differenzierten Messung wurden mit Variationen des NASA Task Load Index (Hart & Staveland, 1988) gestartet. Allerdings wurden die entsprechenden Variationen nicht in den Beiträgen veröffentlicht (Gerjets, Scheiter & Catrambone, 2006; Zumbach & Mohraz, 2008), eine Replikation ist daher leider nicht möglich. Naismith, Cheung, Ringsted und Cavalcanti (2015) setzten ebenfalls den NASA-TLX ein, und verglichen diesen mit einem selbst erstellten „Cognitive Load Component"-Fragebogen mit sechs Items. Es bleibt jedoch unklar, wie die selbsterstellten Items zu Skalen zusammengefasst wurden und somit leider auch, wie diese mit dem NASA-TLX verglichen wurden. In einer weiteren Studie setzten Naismith, Haji et al. (2015) den CLC-Fragebogen erneut ein, allerdings wird auch hier nicht explizit publiziert, wie die Items zu Skalen zusammengefasst werden. Neuere Forschung versucht Abwandlungen des NASA-TLX zu vermeiden und setzte diesen im Original (Aldekhyl, Cavalcanti & Naismith, 2018) oder gekürzt um das Item „Frustration" (Duchowski et al., 2018) ein.

Mehrere Forscher versuchten die Arten von kognitiven Belastung beim Lernen mit einzelnen differenzierten Fragen zu erfassen (z.B. Cierniak, Scheiter & Gerjets, 2009; Cierniak, 2011; Okuni & Widyanti, 2019) und griffen dabei auf einzelnen Items anderer (z.B. Ayres, 2006; Kalyuga et al., 1998; Salomon, 1984) in unterschiedlicher Kombinationen zurück.

Umfangreichere Fragebögen wurden von Swaak und de Jong (2001), Leppink, Paas, van der Vleuten, van Gog und van Merriënboer (2013) und Leppink, Paas, van Gog, van der Vleuten und van Merriënboer (2014) entwickelt. Der Fragebogen von Swaak und de Jong (2001) wurde von Eysink et al. (2009) bereits eingesetzt und erweitert.

Die von Leppink und Kollegen (Leppink et al., 2013, 2014) entwickelten Items wurden mittlerweile bereits mehrfach eingesetzt. Die Items beziehen sich direkt auf Inhalte des Lernstoffes (z.B. Formeln), somit ist eine konkrete Anpassung an den Lerninhalt in der Regel nötig. Beim Einsatz in weiteren Studien bleibt manchmal unklar, ob und wie die Items adaptiert werden, da diese nicht automatisch zu jeder Lernsituation passen (vergleiche Rahimi & Sayyadi, 2019, unklar bleibt hier beispielsweise, wie das Item zu Formeln, bei einer Höraufgabe im nicht mathematischen Kontext verändert oder angepasst wurde).

Aufgrund der teils unbefriedigenden Forschungslage zur differenzierten Messung kognitiver Belastung wurde ein Fragebogen entwickelt der im nächsten Teil der Dissertation durch die eingefügten Beiträge vorgestellt wird. In Kapitel 3 wird die Entwicklung des Fragebogens dargestellt, dabei steht die Verifikation der Fragebogenstruktur im Vordergrund. Die Beiträge in Kapitel 4 überprüfen die Annahmen der CLT-Theorie durch Einsatz des Fragebogens in unterschiedlichen Lernszenarien.

Teil II.

Fragebogen zur differenzierten Messung kognitiver Belastung beim Lernen und dessen praktischer Einsatz im Rahmen von Instruktionsdesignfragestellungen

> Unsere größte Schwäche liegt im Aufgeben. Der sicherste Weg zum Erfolg ist immer, es noch einmal zu versuchen.
>
> *(Thomas Alva Edison)*

3. Fragebogenentwicklung

Ziel der nun eingebundenen Zeitschriften- und Tagungsbeiträge ist, die im vorherigen Abschnitt angesprochenen positiven und negativen Aspekte der Messung kognitiver Belastung beim Lernen bei der Entwicklung eines neuen Fragebogens zur differenzierten Messung kognitiver Belastung beim Lernen zu berücksichtigen.

Der Fragebogen sollte im speziellen

- den instrinsic cognitive Load, den extraneous cognitive Load und den germane cognitive Load, die für Instruktionsdesignfragestellungen wichtigen Arten kognitiver Belastung, differenziert messen und
- er sollte unabhängig vom eigentlichen Lerninhalt und Lernmedium sein, um eine Übertragung auf verschiedenste Lehr- und Lernsettings zu ermöglichen und damit universell einsetzbar sein.

Folgende Publikationen sind daher auf den folgenden Seiten als Volltext in dieser Dissertation eingebunden:

- Klepsch, M., Schmitz, F. & Seufert, T. (2017). Development and Validation of Two Instruments Measuring Intrinsic, Extraneous, and Germane Cognitive Load. *Frontiers in Psychology, 8*, 1997. doi: 10.3389/fpsyg.2017.01997
- Pichler, M. & Seufert, T. (2011). Two strategies to measure Cognitive Load. In European Association for Research on Learning and Instruction (Hg.), *EARLI Conference 2011 Ëducation for a Global Networked Society": Book of Abstracts and Extended Summaries* (S. 928–929). University of Exeter.
- Klepsch, M. & Seufert, T. (9. April 2012). *Subjective Differentiated Measurement of Cognitive Load*. 5th International Cognitive Load Theory Conference, Tallahassee (USA).

Die beiden Tagungsbeiträge von 2011 und 2012 beschreiben jeweils Teile der beiden im Zeitschriftenartikel von 2017 veröffentlichten Studien. Zum Zeitpunkt der Einreichungstermine für die Tagung im Jahr 2011 war gerade die Hälfte der ersten Studie durchgeführt. 2012 war gerade die zweite Studie fertig erhoben und vorläufige Ergebnisse konnten berichtet werden.

Im Zeitschriftenartikel von 2017 wurde alles zusammengefasst und um weitere Analysen ergänzt, weshalb dieser als wichtigster der drei Beiträge an den Anfang gesetzt wurde, während die beiden Tagungsbeiträge lediglich die Entwicklung verdeutlichen.

Klepsch, M., Schmitz, F. & Seufert, T. (2017). Development and Validation of Two Instruments Measuring Intrinsic, Extraneous, and Germane Cognitive Load. *Frontiers in Psychology, 8*, 1997. doi: 10.3389/fpsyg.2017.01997

Abrufbar unter: https://doi.org/10.3389/fpsyg.2017.01997

frontiers in Psychology

ORIGINAL RESEARCH
published: 16 November 2017
doi: 10.3389/fpsyg.2017.01997

Development and Validation of Two Instruments Measuring Intrinsic, Extraneous, and Germane Cognitive Load

Melina Klepsch[1], Florian Schmitz[2] and Tina Seufert[1]*

[1] Department Learning and Instruction, Institute of Psychology and Education, Ulm University, Ulm, Germany, [2] Department Individual Differences and Psychological Assessment, Institute of Psychology and Education, Ulm University, Ulm, Germany

Cognitive Load Theory is one of the most powerful research frameworks in educational research. Beside theoretical discussions about the conceptual parts of cognitive load, the main challenge within this framework is that there is still no measurement instrument for the different aspects of cognitive load, namely intrinsic, extraneous, and germane cognitive load. Hence, the goal of this paper is to develop a differentiated measurement of cognitive load. In Study 1 ($N = 97$), we developed and analyzed two strategies to measure cognitive load in a differentiated way: (1) Informed rating: We trained learners in differentiating the concepts of cognitive load, so that they could rate them in an informed way. They were asked then to rate 24 different learning situations or learning materials related to either high or low intrinsic, extraneous, or germane load. (2) Naïve rating: For this type of rating of cognitive load we developed a questionnaire with two to three items for each type of load. With this questionnaire, the same learning situations had to be rated. In the second study (N = between 65 and 95 for each task), we improved the instrument for the naïve rating. For each study, we analyzed whether the instruments are reliable and valid, for Study 1, we also checked for comparability of the two measurement strategies. In Study 2, we conducted a simultaneous scenario based factor analysis. The informed rating seems to be a promising strategy to assess the different aspects of cognitive load, but it seems not economic and feasible for larger studies and a standardized training would be necessary. The improved version of the naïve rating turned out to be a useful, feasible, and reliable instrument. Ongoing studies analyze the conceptual validity of this measurement with up to now promising results.

Keywords: Cognitive Load Theory, differentiated measurement, instructional design, multimedia research, multimedia design principles

OPEN ACCESS

Edited by:
Douglas F. Kauffman, Boston University School of Medicine, United States

Reviewed by:
Emily Grossnickle Peterson, American University, United States
Ludmila Nunes, Purdue University, United States

****Correspondence:***
Melina Klepsch
melina.klepsch@uni-ulm.de

Specialty section:
This article was submitted to Educational Psychology, a section of the journal Frontiers in Psychology

Received: *12 July 2017*
Accepted: *31 October 2017*
Published: *16 November 2017*

Citation:
Klepsch M, Schmitz F and Seufert T (2017) Development and Validation of Two Instruments Measuring Intrinsic, Extraneous, and Germane Cognitive Load. Front. Psychol. 8:1997. doi: 10.3389/fpsyg.2017.01997

INTRODUCTION AND THEORETICAL BACKGROUND

One of the most powerful and most debated frameworks in educational research during the last few decades has been Cognitive Load Theory (CLT; Sweller and Chandler, 1991; Sweller, 2010a), which is extensively used for evaluating learning environments or interpreting empirical results. However, the absence of an adequate measure of cognitive load has been criticized (Moreno, 2010). The vast majority of studies on multimedia learning assess cognitive load by using a single item to assess perceived invested mental effort (Paas, 1992). Some other studies used objective techniques, such

as dual task measures (e.g., Brünken et al., 2004), or physiological parameters (e.g., heart rate: Paas and van Merriënboer, 1994; eye blink parameters: Goldstein et al., 1992). All these measures show different strengths and weaknesses as will be discussed below. However, until now there has been no instrument that allows measuring the three types of load differentially, namely intrinsic (ICL), extraneous (ECL), and germane cognitive load (GCL) (Brünken et al., 2010; Moreno and Park, 2010). Such a differentiated measure would help to improve our understanding of which aspect of the learning task requires cognitive resources to gain insight into the learning processes and the effects of instructional design.

Cognitive Load Theory (CLT)

In a learning situation, information must be processed in working memory and stored in long-term memory. One of the main assumptions of CLT is that working memory is limited in capacity, when processing information, as well as in time, when it comes to holding information (Paas et al., 2010). The second assumption of CLT is that long-term memory is virtually unlimited (Sweller et al., 1998), and, according to schema theory, knowledge is organized and retained in the form of schemata (Rumelhart, 1981).

The merit of CLT is to make prescriptions for instructional design that reflect these specific characteristics of human cognitive architecture. Following these prescriptions, successful instruction should allow learners to manage working memory load and, hence, to learn successfully. For a long time, CLT differentiated three independent sources of memory load, namely ICL, ECL, and GCL (e.g., Sweller et al., 1998). While intrinsic load arises from the number of interrelated elements of the learning task, extraneous load is caused by the additional requirements that are produced by suboptimal instructional design and are not directly linked to the task. The third source of memory load is germane load, which "reflects the effort that contributes to the construction of schemas" (Sweller et al., 1998, p. 259). All three aspects will be described in the following paragraphs as they inspired the construction of the developed instruments.

Intrinsic cognitive load (Sweller, 1994; Sweller and Chandler, 1994) is the load resulting from the inherent complexity of the learning task. This load type depends on two different factors (Moreno and Park, 2010), the element interactivity of the task and the learner's prior knowledge. (1) *Element interactivity* (Chandler and Sweller, 1996) corresponds to the number of elements that the learner must simultaneously process in working memory while dealing with the task. Low element interactivity, therefore, means that a learner can process the elements sequentially as there is only minimal reference between the elements, whereas tasks with high element interactivity comprise elements that are highly interlinked and must, therefore, be processed simultaneously (Sweller, 2010b). The intrinsic load in language learning is, for example, low when you have to repeat unrelated vocabularies instead of constructing sentences by using the correct grammar. (2) The *prior knowledge* of the learner plays a fundamental role as new information can be linked with existing schemata (Gerjets et al., 2004). Hence, the information that must be processed is more comprehensive and well structured so learners do not have to process (too) many unrelated elements in working memory. A typical example is an expert in chess who can build meaningful chunks instead of memorizing many unrelated elements. According to these two factors, two approaches to manipulate or reduce ICL, have been empirically tested: (1) The *segmenting principle* (Mayer and Moreno, 2010) aims at reducing element interactivity by presenting information step by step, which helps learners with insufficient prior knowledge to organize the incoming information. So with each step, they gain more prior knowledge and are able to better link the incoming information of the next step. The segmentation into steps is a matter of design and can be done by presenting information in parts, for example on different sheets or in the formatting of paragraphs. This describes why segmenting may also reduce ECL. (2) The *pretraining principle* (Mayer, 2005a) reduces ICL by providing the learner with information about the content before starting with the learning material. Increasing the learner's prior knowledge supports the integration of new information. In general, it is difficult to substantially reduce intrinsic load without altering the learning objectives. It is usually not possible to reduce information or complexity without altering the task itself by skipping something. When students have to understand a specific issue, they must face all the relevant ideas and their inherent structure. Therefore, only altering the task itself, e.g., by adding or deleting information or having learners with different levels of prior knowledge, results in a real variation of ICL and not a "virtual" reduction of ICL because of design effects. As one may notice, trying to change ICL can easily end up in also changing ECL. Hence, most instructional designers do not concentrate on what must be learned but on how to present the necessary information and therefore, on ECL.

Extraneous cognitive load is caused by the instructional design of the learning material. Whenever the learner must invest mental resources into processes that are not relevant for the task itself, like searching for or repressing information, we call them extraneous processes. Consequently, ECL is reduced when necessary processes, such as imagery or linking information, is facilitated by design. This means that the designer of the learning material can manipulate ECL in a relatively easy way.

Hence, many researchers addressed the question of how to reduce ECL in a reasonable way. They found vast empirical evidence for such effects, also called multimedia design principles. A few of the principles, which we refer to in the following studies, are included herein. (1) The *multimedia principle* (Mayer, 2001; Fletcher and Tobias, 2005): numerous empirical findings show that learners perform better when learning from a combination of text and pictures rather than from pictures alone. Nevertheless, there are some boundary conditions to this principle, as described in a variety of other principles (e.g., the *redundancy principle*: Sweller and Chandler, 1994 or the *coherence principle*: see below). (2) The *modality principle* (Mayer and Moreno, 2003; Low and Sweller, 2005; Mayer, 2005a): According to this principle it is better to present a text auditorily, not visually, when combined with a picture. While printed text combined with a picture requires visual resources, the auditory text uses the phonological system of working memory (Baddeley, 2003). Since this system is independent

from the visual system, the learner can process text and picture simultaneously, which helps to integrate the two sources. So, working memory capacity is used more effectively, and an overload of the visual capacity is prevented. (3) The *coherence principle* (Mayer and Moreno, 2010): This principle recommends leaving out all decorative information in multimedia learning environments that does not contribute to the learning task itself. For example, decorative pictures may be motivating, but they distract attention and need to be disregarded as they are not relevant for learning (as a connected concept, see also *seductive details effect*: Rey, 2012). (4) The *split attention effect* (Ayres and Sweller, 2005): Split attention occurs when learners are forced to mentally integrate separated sources of information that could be better presented in an integrated way. This effect is also known as the *spatial contiguity principle* (Mayer and Moreno, 2010), which occurs when learners benefit from a spatial integration, e.g., when text information is presented within instead of beside a picture. The integration could also take place timely (*temporal contiguity principle*: Mayer and Moreno, 2010), so text information accompanying dynamic pictures like video or animation should be presented simultaneously instead of delayed. The integrated, or simultaneous, presentation facilitates synchronous mental processing and, hence, the integration of information.

Overall, there are many ways to optimize instructional design to reduce ECL (for more principles/effects, see Mayer, 2005b; Sweller, 2010a). With the freed-up resources, learners could invest more effort into in-depth learning processes.

Germane cognitive load, the third type of cognitive load, results from activities required of a learner that facilitate learning and contribute to transfer performance, helping to build correct mental models (Paas et al., 2003a). Such activities could be, for example, taking notes during reading a text, coming up with memory hooks to remember something, or explaining the learned content to someone else. Therefore, high germane load indicates that learners are engaged and direct their mental resources to learning processes. Researchers also addressed the question of how to foster this type of load (e.g., Paas and van Gog, 2006) and described several design principles. In the following studies, we refer to the *self-explanation effect* (VanLehn et al., 1992) for varying GCL. This effect indicates that learners who explain learning material to themselves achieve a higher learning outcome. They actively link the new information with existing schemata and, hence, invest in deeper learning processes. In our studies, we used different approaches to induce self-explanation. We, for example, asked learners to engage in different learning activities, e.g., to formulate titles to text paragraphs, to find examples to a topic, or to sum up a topic in their own words. Other strategies have been to engage memory strategies through prompts (Berthold et al., 2007, 2009).

The given descriptions make it obvious that germane load is highly dependent on intrinsic load. Moreover, learners are only able to devote germane resources if the amount of extraneous load is not exceeding their working memory capacity. Hence, germane load is also linked to extraneous load. Therefore, an attempt was made to refine CLT and conceptually differentiate those aspects of load that are either intrinsic to the task, and therefore productive, or those that are extraneous, and therefore unproductive. Hence, a new definition of germane load is that it "refers to the working memory resources available to deal with the element interactivity associated with ICL" (Sweller, 2010b, p. 126). From a theoretical point of view, this clarification is helpful and necessary. From a measurement point of view, it is nevertheless relevant to understand all loading aspects in a learning situation: the resources required by the task or those resources available that are deliberately devoted by the learner. Therefore, we address all aspects of cognitive load: the intrinsic and productive aspects, including the given intrinsic element interactivity and the devoted germane resources, to understand these interactions and the extraneous and unproductive load. Given the three-partite nature of cognitive load, *ICL*, *ECL,* and *GCL* all need to be considered, as all of them are important and noteworthy when generating and designing learning material. However, to date, there has been no instrument that allows for the assessment of all three types of load in a differentiated way.

Cognitive Load Measurement

Many researchers stated that measuring cognitive load is one of the persistent challenges in educational research (Mayer and Moreno, 2002; Brünken et al., 2003, 2010; Schnotz and Kürschner, 2007; de Jong, 2010; Moreno, 2010). It has even been questioned whether it is a "mission impossible" (Brünken et al., 2010) facing the measurement approaches at hand. Those approaches that use either subjective ratings or objective measures, such as dual-task performance, are directly addressing load, e.g., by asking learners to rate their perceived mental load, or indirectly by using indicators that are connected to load, such as performance measures. Moreover, they are either online and measure constantly or repeatedly during the learning process or offline and measure retrospectively. All these measures show strengths and weaknesses (see Brünken et al., 2003, 2010). In the following pages, we will describe and briefly discuss the most common measurement approaches: (1) self-report measures, (2) dual-task measures, and (3) measures of physiological parameters.

Self-report Measures

The most popular scale for measuring cognitive load is a rating scale developed by Paas (1992). This scale is a modified version of a scale by Bratfisch et al. (1972), which was constructed to measure task difficulty. In fact, the scale used by Paas consists of one item, and participants are asked to use a 9-point Likert scale for their responses, ranging from very, very low mental effort (1) to very, very high mental effort (9). The exact wording of the item is not published in the article, but it is similar to ratings used by many other researchers (for an overview, see Paas et al., 2003b). Typical item wordings: "I invested . . . mental effort" or "my invested mental effort was . . .." The scale is usually designed as a 5- to 9-point Likert scale. While one-item scales appear to be economic at first glance, they are generally problematic from a psychometric perspective. There is no way to tease apart true variance from measurement error. This is a problem discussed in short by van Gog and Paas (2008). They complain about different wordings, different labels, and an inconsistent scale range. The

fact that introspective ratings are not highly reliable has been demonstrated by the study of van Gog et al. (2011), where overall retrospective measures of load were generally higher than the mean of several measures during learning. Learners adjust their ratings with respect to situational parameters, and they use subjective internal standards to evaluate their current load state, if they even possess the ability of introspection.

More importantly, one item cannot distinguish between the various sources of load postulated in CLT. In fact, sources of load have been inferred from the combination of task performance with rated mental effort in some previous studies: If learning outcome is low, despite highly rated mental effort, this is interpreted as resulting from inappropriate affordances of the learning materials (i.e., extraneous load). But, high task difficulty (intrinsic load) offers a plausible explanation too. However, such inferences from outcome likely result in a circular argument (Kalyuga, 2011; Kirschner et al., 2011) and should be avoided.

A second frequently used measurement technique is a one-item rating of task difficulty (DeLeeuw and Mayer, 2008; for an overview, see van Gog and Paas, 2008 or Brünken et al., 2010). Overall, its limitations are the same as for the mental-effort rating. For a broader overview on studies using one-item subjective cognitive load measures, also see Paas et al. (2003b).

For self-report measures that measure different aspects of CLT see section "Measuring Cognitive Load in a Differentiated Way".

Dual-Task Measures

Cognitive load measures using the dual-task paradigm require a learner to perform two tasks simultaneously. It is assumed that performance for the second task drops when the primary task, i.e., the learning task, becomes more loading. There are two possible ways to conduct dual-task measures. (1) On the one hand, it is possible to measure accuracy and response times in an observation task that needs to be carried out during the performance of the learning task (e.g., Brünken et al., 2002). (2) On the other hand, a concurrent second task needs to be performed during learning (e.g., Park and Brünken, 2011). This, for example, could be tapping your feet in a given rhythm as constant as possible. Increasing load in the first task could then be measured by impairments of the secondary tasks. Dual-task measures of load have the advantage of being objective, and they mirror the whole learning process so you can gather rich data. However, the most obvious disadvantage is the intrusiveness of such techniques; they disturb the learning process and impose load by themselves. Besides it is a matter of resources: Learners with high working memory capacity might not be as loaded by a secondary task as learners with low working memory capacity. This would always result in the need to control the prerequisites of learners associated with working memory. Another disadvantage is, as already mentioned for the subjective ratings, that it is also not possible to identify the type of load that is measured.

Measures of Physiological Parameters

A wide range of physiological parameters have been used as indicators for cognitive load. The most commonly used physiological parameters are heart rate (Paas and van Merriënboer, 1994), pupil dilation (van Gerven et al., 2004), and electroencephalography measures (Antonenko et al., 2010). There are also some less common ones, such as measuring hormone levels (Wilson and Eggemeier, 1991) or using fMRI measures (Whelan, 2007). However, it is difficult to tell what triggered the physiological processes and, hence, to interpret the data (Brünken et al., 2010). Moreover, the measures are usually intrusive and less economic, and the problem is again that it is impossible to tell if ICL, ECL, or GCL is being measured.

Measuring Cognitive Load in a Differentiated Way

A major shortcoming of the previously discussed approaches is that they only assess the overall amount of experienced load and do not distinguish between intrinsic, germane, or extraneous load. This limits their usefulness in instructional design and multimedia learning research that build on the differentiated CLT. To overcome this limitation, some researchers have developed questionnaires to measure cognitive load in a differentiated way.

Several researchers (e.g., Gerjets et al., 2006; Zumbach and Mohraz, 2008) used variations of the NASA Task Load Index (Hart and Staveland, 1988) in an attempt to measure cognitive load in a differentiated way. However, the wordings of those variations are not always well documented. Also, the study of Naismith et al. (2015), in which they compare their questionnaire called cognitive load component (CLC) with the NASA-TLX also does not provide sufficient information on the items and scales.

Cierniak et al. (2009) and Cierniak (2011) used three items to represent ICL, ECL, and GCL. In their studies, they found significant matches between performance data and measured cognitive load. They asked for "difficulty of the learning content" (adopted from Ayres, 2006) to represent ICL, "difficulty to learn with the material" (adopted from Kalyuga et al., 1998) to represent ECL, and "concentration during learning" (adopted from Salomon, 1984) to represent GCL. However, they didn't test their questions with a wide variety of learning material to validate their scale.

Swaak and de Jong (2001) used a measurement scale called SOS to measures participants' cognitive load during learning in a simulation environment about electrical circuits. In a first version, they used three items to measure difficulty of the subject, difficulty of working with the operating system, and usability of support tools. It was not explicitly specified which item refers to which type of load. All three items had to be answered on a scale ranging from "extremely easy" (0) to "extremely difficult" (100). They did not find differences in cognitive load in their study between the experimental groups. An adapted and extended version of the SOS scale was used by Eysink et al. (2009) in multimedia learning arrangements. ICL was measured by asking for the perceived difficulty of the domain. ECL was measured using three items asking for accessibility of information (collect all needed information), design (distinguishing important from unimportant information), and navigation (working with the learning environment). GCL was measured by asking for

the difficulty of understanding the simulation. They found differences for all three types of cognitive load, but they were not theoretically linked or discussed regarding the learning outcome, i.e., productive or unproductive aspects.

One of the most recent approaches to measure cognitive load differentially that has attracted attention in the field is a scale by Leppink et al. (2013). They developed a questionnaire for complex knowledge domains consisting of 10 items, which were tested in the domain of statistics. The questionnaire included three items on ICL that asked about the complexity of the topics, formulas, and concepts of the activity. It included three items on ECL that refer to the instruction and/or explanations given during the activity and asked, e.g., about ineffectivity. For GCL four items were included that referred to enhancement of understanding of the topics, formulas, concepts, and definitions. One of the GCL items particularly referred to the understanding of statistics. According to the authors, the term *statistics* could be replaced by any other knowledge domain. They found promising results and tried to replicate their findings in another set of experiments (Leppink et al., 2014). In Study 1, they replicated the factor structure in a similar context (statistics education) but within a different context (language learning). Even if they had to adopt the items to fit the different domains, they found the three factors to be robust. In Study 2, they again adopted the items to fit the domain and added three more items explicitly asking for "invested mental effort" on the three types of load. In their studies, they were missing a positive correlation between items that are supposed to measure GCL and learning outcomes as along with a substantial correlation between the "old" items and the new one. As a conclusion, according to Sweller et al. (2011), they only approved the measurement of intrinsic and extraneous load as being meaningful.

Based on these previous attempts to measure cognitive load differentially, we developed and evaluated self-report measures that tap the different aspects of cognitive load in a differentiated way. We decided on a self-reported measure of load due to its economy and flexibility. Our goal was to develop a reliable domain-unspecific questionnaire that could be validated and used in various learning situations. Whereas measures of ECL and ICL within our questionnaire should evaluate the inherent complexity and the design of the learning material as it was perceived by the learner (based on his prior knowledge and expertise and his prerequisites for different instructional designs) during learning, measures of GCL should focus on the additional investment of cognitive processes into learning (triggered through elements that learning material contains, e.g., prompts). The most straightforward way to address this challenge was to develop a self-report questionnaire—which we will refer to as naïve rating in the following. Items would reflect all three aspects of cognitive load. We additionally followed another path, in which we qualified students to understand and differentiate the three aspects of cognitive load—we will refer to this as an informed rating in the future. The questionnaire in this case asked, without paraphrasing, to rate the intrinsic, extraneous, and germane load for each learning scenario.

To analyze whether the two approaches to measure cognitive load in a differentiated way are reliable, valid, and comparable, we conducted two experimental studies. In our first study, we were interested in the comparability and quality of the two instruments. In the second study—with an additional pilot-study—we focused on the refinement of the questionnaire for the naïve rating and, hence, on developing a new, economic, domain-independent, and especially differentiated measure of cognitive load.

STUDY 1

This first study was conducted to compare the two approaches to measure cognitive load, the informed rating and the questionnaire for a rating without prior knowledge about the concepts of load, which we, therefore, call a naïve rating. We could have used a learning scenario in which learners carry out a learning task and rate their experienced load in this situation afterward, as most other researchers have chosen to do (e.g., Cierniak et al., 2009; Leppink et al., 2013). To evaluate whether our instruments could detect differences—in the amount of load and in the type of load—in multiple different learning situations, we used descriptions of hypothetical learning situations (verbally or in pictorial form via screenshots). Each scenario had to be evaluated by means of the respective questionnaires. Therefore, learning situations and load rating were hypothetical. However, one strength of this approach was that it allowed us to cross-validate our measures in various settings and to compare our two instruments for several ratings. Additionally, we analyzed the reliability of our scales.

Methods and Materials

Participants and Design

The participants in Study 1 included 97 computer science or psychology students from a German university in their first or third semester. No other demographic data were assessed. Each participant was randomly assigned to the informed rating group ($n = 48$) or the naïve rating group ($n = 49$). Dependent variables were ICL, ECL, and GCL for each task, either measured by the informed or naïve questionnaire.

Procedure

At the beginning, all participants were informed about the procedure and signed an informed consent, and participants were aware that they could withdraw their data at any point in the study. Afterward, each participant was randomly assigned to one of the two groups: (1) The informed rating group, which got an introduction to CLT, as described later, to be able to rate the following tasks in an informed way, and (2) the naïve rating group, whose members did not get any previous information about CLT but started directly with the evaluation tasks. The evaluation tasks, as described below, were small learning tasks or scenarios which had to be conducted. Participants also were asked to write down the correct answer, if possible. After each task, each participant filled out either the informed rating questionnaire or the naïve rating questionnaire, both described below. Altogether, the study took about 90 min for the informed rating group and 60 min for the naïve rating group.

Evaluation Tasks

For this experimental study, we developed 24 learning tasks or scenarios grouped in five different domains (language learning, biology, mathematics, technology, and didactics). In Study 1, tasks within each domain varied in only one aspect of cognitive load (ICL, ECL, or GCL), following from theoretical accounts of CLT, empirical results of cognitive load related studies, or the above-mentioned multimedia design principles. All evaluation tasks are displayed in Supplementary Table S1. It should be mentioned that we quickly noticed that the implementation of a variation of one type of load did not affect another type of load. Learners had to rate all three types of cognitive load for all these tasks/scenarios. Here are some examples of variation: (1) For ICL, learners had to rate cognitive load for a task where the element-interactivity had been varied, e.g., "the day after tomorrow will be Saturday. Which day was yesterday?" which was supposed to be rated with lower ICL scores versus "3 days after yesterday was Friday. Which day will be 5 days before tomorrow?" where the ICL should be rated higher. In the vocabulary tasks, we used, for example, languages with different inherent complexity. (2) For extraneous load, we showed learning environments with an integrated format of text and picture (lower ECL) versus a separated format (higher ECL). Another variation we used was material with (high extraneous load) or without seductive details (low extraneous load). (3) For germane load, we asked participants to rate different instructional settings that should either induce GCL, like "every 20 min a teacher gives you time to think of examples you can find for the topic" versus tasks without such an activation. Nevertheless, learners had to rate all three types of load after conducting each task to examine whether the ratings in fact only differed with respect to the theoretically assumed type of load. Everyone was also asked to write down the correct answer for the given learning task. For the learning scenarios used for variations of GCL which had to be merely imagined, nothing had to be written down.

Cognitive Load Measures

Informed rating

If learners know what these load types are and in which way they differ, they should be able to rate the three types of cognitive load correctly. Thus, we first developed an introduction into CLT. The introduction included information about CLT itself, working memory, and types of load. It was presented as a lecture using PowerPoint slides. The lecture ended with a few examples, showing how variations of the three types of load could be implemented in typical learning materials. These examples were different from the tasks to be rated. The design principles and domains used were sometimes overlapping, but the tasks themselves were completely new to the participants to prevent replicating information from the lecture. The aim was to qualify the participants to detect the three types of load and in which ways they are interrelated and how they can be differentiated from each other. The whole training, including a discussion, lasted about 30 min. Afterward the training, the participants might not be experts, but they should be able to rate cognitive load differentially in an informed way. After the training, we handed out a written summary of the three types of cognitive load. This little booklet was allowed during rating, so our participants were able to look up the types of load during rating if they did not feel confident with the concepts of load. The developed questionnaire for informed rating directly targets the three types of load. Therefore, we developed three items, which read as follows: (1) "During this task, ICL was. . .," (2) "During this task, ECL was. . .," and (3) "During this task, GCL was. . .." As a fourth question, the informed rating questionnaire included a question about the overall mental load during the learning situation that was adopted from Paas (1992). All items had to be rated on a 7-point Likert scale from "very low" to "very high."

Naïve rating

The other group of participants was not informed about the concept of cognitive load and, therefore, rated the same learning situations in a naïve way by completing the self-report questionnaire. This first version comprised two questions related to ICL, three related to ECL, and another two items related to GCL. All items had to be rated on 7-point Likert scales from "completely wrong" to "absolutely right." As an eighth question, the naïve rating included the same question as the informed rating about the overall mental load during the learning situation. All items for the naïve rating are shown in **Table 1**. The items of the questionnaire were developed in German and were translated into English for this paper. An advantage of this approach is that it does not require an introduction into CLT. However, the aim was to find items that are very clear, so participants would easily understand the questions; also, the questions needed to be clearly related with the respective sources of cognitive load.

Data Analysis

Reliability was analyzed separately for both instruments. To this end, we calculated internal consistency for each of the 24 tasks. For the *informed rating*, there was only one item per load type. Consequently, we analyzed whether the three load types form one single construct, or—as we would expect—three separable constructs, i.e., internal consistency of all three items together should be rather low. To report this data, a meta-analysis of coefficient alpha based on formulas presented by Rodriguez and Maeda (2006) has been conducted, which allows us to conduct a mean of several given alphas based on sampling distribution. For the *naïve ratings*, there were several items for each load type, and internal consistency could be computed for the items of each scale (separately for each task), and this was predicted to be high. Again, to report this data in an aggregated way, the before-mentioned weighted mean was conducted to generalize Cronbach's alpha. For the naïve rating, we also calculated internal consistency for all seven items and aggregated them.

Validity was analyzed by comparing the learners' ratings with the theoretically predicted outcome, i.e., whether the participants rated the specific loads as either low or high, according to our intended task design. Therefore, tasks with differing load levels should be rated significantly different in the amount of this specific load. This means, e.g., all tasks developed to induce low intrinsic load should be rated significantly lower than all tasks developed to induce high intrinsic load.

TABLE 1 | Items of the first version of the naïve rating questionnaire.

Type of load	Item - German	Item - English
ICL	Bei der Aufgabe musste man viele Dinge gleichzeitig im Kopf bearbeiten.	For this task, many things needed to be kept in mind simultaneously.
ICL	Diese Aufgabe war sehr komplex.	This task was very complex.
GCL	Bei dieser Aufgabe musste ich selbst ganz aktiv nachdenken.	For this task, I had to highly engage myself.
GCL	Bei dieser Aufgabe musste ich intensiv überlegen, wie einzelne Dinge gemeint sind.	For this task, I had to think intensively what things meant.
ECL	Bei dieser Aufgabe ist es mühsam, die wichtigsten Informationen zu erkennen.	During this task, it was exhausting to find the important information.
ECL	Die Darstellung bei dieser Aufgabe ist ungünstig, um wirklich etwas zu lernen.	The design of this task was very inconvenient for learning.
ECL	Bei dieser Aufgabe ist es schwer, die zentralen Inhalte miteinander in Verbindung zu bringen.	During this task, it was difficult to recognize and link the crucial information.

ICL, intrinsic cognitive load; ECL, extraneous cognitive load; GCL, germane cognitive load.

To *compare the two instruments*, we compared means by a *t*-test. This should reveal, whether there are significant differences between the scales for ICL, ECL, and GCL of the informed and naïve ratings. We expected the two ratings not to differ and followed a conservative approach, when we decided that we accept the ratings not to be significantly different if $p > 0.20$[1].

Comparability and validity will be analyzed by means of a mixed analysis of variance (ANOVA) for each type of load.

Finally, relations with the established global load measure by Paas (1992) were investigated. Each participant in the informed and the naïve rating group also filled out this scale for each task. For the informed rating, we correlated the three items separately for the mental effort item. For the naïve rating, we correlated our three load-type scales for the mental effort item. We predicted moderate to high correlations, as mental effort should be rated high whenever a task implies a high load—irrespective of which load type—and should be low whenever a task is designed to result in low load. Therefore, for both ratings we also report correlations of the sum of the three load ratings with the mental effort rating.

Results

The *reliability* of the informed rating test depended on tasks. For each task, we analyzed the internal consistency of the ICL, ECL, and GCL item. As expected, reliability was low for all the tasks. Reliability generalization of Cronbach's α resulted in a low $\alpha = 0.25$, as expected due to the different aspects of load which we measured with the three items.

For the naïve rating, we first conducted the reliability test for the two, respectively, three items for each load type for each task. The aggregated alphas that we obtained though reliability generalization have been as expected: For ICL $\alpha = 0.86$, for ECL $\alpha = 0.80$, and for GCL $\alpha = 0.80$. The internal consistency between all ICL, ECL, and GCL items (seven items) for each task in the naïve rating showed an aggregated α of 0.86, which was way higher than expected.

A mixed ANOVA for each type of load was used to *compare the two questionnaires* (naïve vs. informed rating – between subjects) and to *validate* if the rated amount of load between high and low load tasks differs for both questionnaires in general (low vs. high load tasks – within subjects). We also calculated contrasts for each questionnaire between high and low load tasks and contrasts for the differing amount of load between the two instruments.

For ICL (see **Figure 1A** and **Table 2**), we found a main effect for the amount of load [$F(1,89) = 490.23$, $p < 0.001$, $\eta^2 = 0.85$]. Theoretically low ICL tasks ($M = 1.80$, $SD = 0.64$) have been rated significantly lower than theoretically high ICL tasks ($M = 4.29$, $SD = 1.09$) with both instruments. Concerning the validity of the two instruments, respectively, we also found a significant difference between low and high load ratings for each instrument [informed rating: $F(1,89) = 293.32$, $p < 0.001$, $\eta^2 = 0.77$; naïve rating: $F(1,89) = 199.96$, $p < 0.001$, $\eta^2 = 0.69$]. For comparability, no main effect, i.e., no significant difference between the two approaches, has been found [$F(1,89) = 1.18$, $p = 0.28$, $\eta^2 = 0.01$]. Contrasts support this result [low ICL tasks: $F(1,89) = 0.79$, $p = 0.38$, $\eta^2 = 0.01$; high ICL tasks: $F(1,89) = 3.81$, $p = 0.05$, $\eta^2 = 0.04$]. An interaction effect could be found [$F(1,89) = 6.13$, $p < 0.05$, $\eta^2 = 0.06$] because the informed raters rated more extreme than the naïve raters.

For ECL (see **Figure 1B** and **Table 2**), we also found a main effect for the amount of load [$F(1,95) = 275.91$, $p < 0.001$, $\eta^2 = 0.74$]. Theoretically low ECL tasks ($M = 2.91$, $SD = 1.02$) have been rated significantly lower than theoretically high ECL tasks ($M = 4.69$, $SD = 0.96$) across instruments. Contrasts reveal the same results for each questionnaire [informed rating: $F(1,95) = 174.60$, $p < 0.001$, $\eta^2 = 0.65$; naïve rating: $F(1,95) = 105.32$, $p < 0.001$, $\eta^2 = 0.53$]. For comparability, a main effect has been found [$F(1,95) = 8.32$, $p < 0.05$, $\eta^2 = 0.08$]: the informed raters reported higher ECL ratings (low ECL tasks: $M = 3.03$, $SD = 1.06$; high ECL tasks: $M = 5.05$, $SD = 0.86$) than the naïve raters (low ECL tasks: $M = 2.79$, $SD = 0.98$; high ECL tasks: $M = 4.34$, $SD = 0.93$). Contrasts revealed no significantly different ratings between the questionnaires for theoretically low ECL tasks [$F(1,95) = 1.30$, $p = 0.26$, $\eta^2 = 0.01$] but

[1] As the logic of classical significance testing is inverted when testing for the absence of differences, we adopted a more conservative criterion. Thereby, we sought to reduce a possible beta error, i.e., assuming no differences despite there are some in reality.

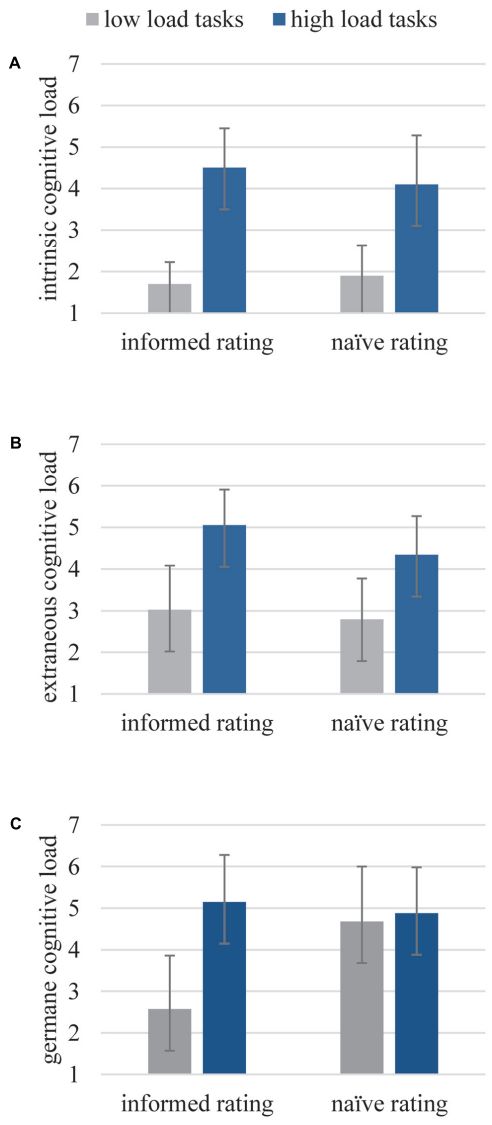

FIGURE 1 | Subjective ratings of the informed and naïve raters for ICL **(A)**, ECL **(B)**, and GCL **(C)** for corresponding theoretical low or high load tasks.

different ratings for theoretically high ECL tasks [$F(1,95) = 14.92$, $p < 0.001$, $\eta^2 = 0.14$]. The interaction effect is also significant [$F(1,95) = 4.72$, $p < 0.05$, $\eta^2 = 0.05$].

For GCL (see **Figure 1C** and **Table 2**), we again found a main effect for the amount of load [$F(1,88) = 53.24$, $p < 0.001$, $\eta^2 = 0.38$]. Again, in general, theoretically low GCL load tasks ($M = 3.67$, $SD = 1.68$) have been rated significantly lower than theoretically high GCL tasks ($M = 4.88$, $SD = 1.13$). The same pattern could be found for the informed rating [$F(1,88) = 105.84$, $p < 0.001$, $\eta^2 = 0.55$]. For the naïve rating, no significant difference between theoretically high and low GCL tasks was found [$F(1,88) = 0.04$, $p = 0.84$, $\eta^2 < 0.001$]. For comparability, we found a main effect [$F(1,88) = 18.02$, $p < 0.001$, $\eta^2 = 0.17$]. The rating of theoretically low load tasks differed between questionnaires [$F(1,88) = 58.66$, $p < 0.001$ $\eta^2 = 0.40$] as did the rating for theoretically high load tasks [$F(1,88) = 4.82$, $p < 0.05$, $\eta^2 = 0.05$]. The informed group rated much more sophisticated (low GCL tasks: $M = 2.57$, $SD = 1.29$; high GCL tasks: $M = 5.14$, $SD = 1.13$), whereas the naïve ratings are on a high level for low GCL tasks ($M = 4.68$, $SD = 1.32$) and high GCL tasks ($M = 4.88$, $SD = 1.10$). Again, a significant interaction effect can be found [$F(1,88) = 57.35$, $p < 0.001$, $\eta^2 = 0.40$].

Concerning the question of whether the differentiated items (informed rating), respectively, scales (naïve rating) fit to ratings on the overall mental-effort item correlations for each task were conducted. Also, the sum of the items/scales and their correlation with the global load measure by Paas (1992) were calculated.

We found substantial correlations for almost all ratings of the informed group: For tasks where ICL was varied, we found correlations between the intrinsic-load rating and the mental-effort item of r between 0.41 and 0.68 for the respective tasks. If ECL was varied, we found correlations between the extraneous-load rating and the mental-effort item of r between 0.32 and 0.60. The germane-load rating correlated with the mental-effort item with r between 0.32 and 0.41 in those tasks where GLC has been varied. When adding together the ratings for ICL, ECL, and GCL of the informed raters as a sum and correlating it with the mental-effort item for each task, we got correlations of r between 0.32 and 0.69.

For the naïve rating concerning tasks where ICL was varied the correlations of the intrinsic-load rating with the mental-effort item resulted in r between 0.47 and 0.81. The correlations of the extraneous-load rating with the mental-effort item was r between 0.36 and 0.68 for tasks where ECL was varied. When GCL was varied, we found correlations of the germane-load rating and the mental-effort item of r between 0.51 and 0.77. Again, when adding together the ratings for ICL, ECL, and GCL of the naïve raters as a sum and correlating it with the mental effort item for each task, we got correlations of r between 0.46 and 0.85.

Discussion

We analyzed and compared two different approaches to measure the three conceptual parts of cognitive load differentially concerning their validity, i.e., their power to confirm theoretically predicted ratings and their reliability.

Our results first provide evidence that the *informed rating* seems to be a valid method of measuring cognitive load in a differentiated way. Participants were able to rate the amount of different aspects of cognitive load as intended by the design and in line with the theoretical predictions. The fact that the overall reliability was very low suggests that learners perceive different levels of load for the three distinct parts of load; ICL, ECL, and GCL. This is not surprising, given that the

TABLE 2 | Means (M) and standard deviations (SD) for subjective ratings of the informed and naïve raters for ICL, ECL, and GCL for corresponding theoretical low or high load tasks.

Type of load		Informed raters		Naïve raters	
		Low load tasks	High load tasks	Low load tasks	High load tasks
ICL	*M (SD)*	1.74 (0.53)	4.52 (0.95)	1.86 (0.73)	4.08 (1.18)
ECL	*M (SD)*	3.02 (1.06)	5.05 (0.86)	2.79 (0.98)	4.34 (0.93)
GCL	*M (SD)*	2.57 (1.29)	5.15 (1.13)	4.68 (1.32)	4.88 (1.10)

ICL, intrinsic cognitive load; ECL, extraneous cognitive load; GCL, germane cognitive load.

three types of load are clearly distinguished in theory (Moreno and Park, 2010). Nevertheless, we should analyze reliability of this measure with other reliability parameters, such as retest-reliability or split half-tests in future studies. Furthermore, as the learners could differentiate between the three types of load and as the independence of the three types seem to be sufficiently given, we argue that an overall measure without differentiating these parts, like the often-used mental-effort scale, is not adequate. To further improve our informed rating, better standardization of the instruction may be indicated. In our case, it was a spoken lecture. A more standardized version of the introduction could instead be given independently from a personal instructor in either written form with a booklet or an interactive e-learning environment. This learning environment should again address all concepts of CLT and contrast learning material with different parameter values of ICL, ECL, and GCL.

The *naïve rating* instrument showed satisfying internal consistency for all three aspects of cognitive load. The scales were also valid for ICL and ECL, but not for GCL. While informed ratings scored germane load of the task as induced, this was not the case for the naïve rating. Learners failed to differentiate between theoretically predicted low and high levels of germane load. Especially problematic was that the naïve ratings for theoretically low GCL tasks were extremely high. Therefore, the current items should be revised to increase the validity of the germane-load scale. It could be speculated that the wording of the current items was ambiguous so learners understood them differently. We should also keep in mind that the informed raters could detect differences in load levels more clearly. The questionnaire also allowed us to differentiate between high and low load levels, albeit it appeared not as sensitive in the present form.

In this study, we designed tasks that should differ in the level of one type of load specifically. However, we cannot design tasks that mirror an exact amount of a special load type. Hence, we cannot qualify the absolute accuracy of the learners' ratings. Moreover, in a naturalistic learning setting all three types of load can always be qualified as either high or low. In future research, real learning tasks should be employed instead of hypothetical ones. Second, all three types of load should be rated for each task based on theoretical assumptions. Finally, larger samples can be recommended for the development of a questionnaire.

Overall, the measure using informed ratings worked arguably well. However, this strategy is not very economic, and, therefore, difficult to put into practice in educational research. Hence, in our second study, we focused on improving the naïve ratings measure. Given that the items for ICL and ECL functioned well, we focused on improving the measurement of GCL and to analyze again the validity and reliability of our questionnaire.

STUDY 2

Based on the results of the first study, we adhered to the approach to measure cognitive load differentially with a questionnaire that can be answered by naïve raters without explicit knowledge of the concepts of CLT. In the second study, our aim was to develop and evaluate a questionnaire that should be able to differentiate between all three aspects of cognitive load. As the germane-load items in the previous questionnaire showed unsatisfying results for validity, we concentrated on redesigning these items first in a pilot study and afterward evaluated the overall questionnaire again in our second study, which was conducted as an online study to facilitate collecting a larger sample.

Pilot Study for Redesigning the Germane-Load Items

To develop new, valid, and reliable items to specifically measure GCL, we conducted a pilot study where nine newly developed items for germane load were generated and tested as described in Study 1. We designed eight tasks with learning scenarios that can be expected to either result in high germane load or not. As an example, to induce GCL, we used prompts to activate learning strategies, such as asking them to write a summary of the given text or prompting them to produce a memory hook. After conducting each task, learners rated the nine new items on a 7-point-Likert-scale (from absolutely wrong to absolutely right). Twenty-seven participants took part in the pilot.

Validity was tested by inspecting which of the newly generated items could discriminate between low and high load tasks. A *t*-test was conducted to this end. This revealed that six out of nine items differentiated well between high and low germane-load tasks (d between 0.36 and 1.68 for each item). As we wanted at least two, but at most only three, items to be in the final version of the questionnaire, we decided to pick items that reflected a wide range of possible influences on GCL, that are suitable for a large variety of learning situations, and have a sufficient effect size: (1) "I made an effort, not only to understand several details, but to understand the overall context," which reflects understanding of the overall context [$t(26) = 4.75$, $p < 0.001$, $d = 0.94$]. (2) "My point while dealing with the

task was to understand everything correctly" to reflect effort of understanding everything correctly [$t(26) = 5.31$, $p < 0.001$, $d = 0.88$]. (3) "The learning task consisted of elements supporting my comprehension of the task" to reflect stimuli for deeper processing by supporting elements within the learning material [$t(26) = 6.17$, $p < 0.001$, $d = 1.68$]. The last item, which targeted supporting elements, is important for studies using worked examples, prompts for learning strategies, or similar elements that should enhance germane load. Consequently, this item may not be fitting for each learning situation but is rather important if germane load is varied on purpose. Original wording of the items in German can be found in **Table 3**.

The final questionnaire for Study 2 comprised the two items for ICL and the three items for ECL already used in Study 1. Additionally, the three novel items for GCL were included. All used items (German wording and English translations) are presented in **Table 3**.

Methods and Materials

Learning Tasks

For this experimental study, we created different learning tasks, which were presented online. Unlike in Study 1 (where each task was designed to vary or induce only one cognitive load type), each task varied ICL, ECL, and GCL to simulate more realistic learning situations. The variations were instantiated as follows: (a) For ICL, we varied element interactivity of a task. (b) For ECL, we presented, for example, learning environments with an integrated format of text and picture versus a separated format or just added non-relevant information. (c) For GCL, we showed learning tasks, which should either induce germane load by activating deeper learning processes versus tasks without such an activation. Experts in a pilot test have validated this classification: They had to rate for each task whether the three types of load would be high or low. Additionally, they were asked if the tasks are useful as evaluation tasks and should end up in the study or not (2 rater, Krippendorf's $\alpha = 0.91$). In the end, we decided on 17 different tasks, belonging to five different learning or problem-solving domains (vocabulary, biography, figure matching, biology, programming). For each topic, two to five tasks, varying ICL, ECL, and GCL (theoretically low versus high), were used (see Supplementary Figure S1 for details). In the vocabulary tasks, as an example, learners always had to learn three Swedish words. In one task, we used words that were similar to their German counterpart (low ICL), included unnecessary information and formatting (high ECL), but no activation was included (low GCL). In another task we used difficult words (high ICL), included memory hooks for two words and asked learners to come up with a memory hook for the third word (high GCL), but did not include unnecessary information or formatting (low ECL).

Participants

Between 65 and 95 participants completed each learning task and rated it. All of them were students of a German university with a major in psychology or computer science in their first or second semester. No other demographic data has been assessed.

Procedure

At the beginning, all participants were informed about the procedure and signed an informed consent through an online form. As participation was voluntary, participants had the chance to withdraw their data at any point in the study. All learning tasks where presented online; learners could, therefore, participate in this study without coming to a lab. After each task, learners posted their answers to the given question or their solution to the given problem and rated their perceived cognitive load with two items on ICL, three items on ECL, and three items on GCL. All items had to be answered on a 7-point Likert scale (ranging from absolutely wrong to absolutely right). The learning tasks were presented in partly random order. For some tasks, it was necessary to keep a specific order; otherwise the previous tasks

TABLE 3 | Items of the second version of the naïve rating questionnaire.

Type of load	Item - German	Item - English
ICL	Bei der Aufgabe musste man viele Dinge gleichzeitig im Kopf bearbeiten.	For this task, many things needed to be kept in mind simultaneously.
ICL	Diese Aufgabe war sehr komplex.	This task was very complex.
GCL	Ich habe mich angestrengt, mir nicht nur einzelne Dinge zu merken, sondern auch den Gesamtzusammenhang zu verstehen.	I made an effort, not only to understand several details, but to understand the overall context.
GCL	Es ging mir beim Bearbeiten der Lerneinheit darum, alles richtig zu verstehen.	My point while dealing with the task was to understand everything correct.
GCL*	Die Lerneinheit enthielt Elemente, die mich unterstützten, den Lernstoff besser zu verstehen.	The learning task consisted of elements supporting my comprehension of the task.
ECL	Bei dieser Aufgabe ist es mühsam, die wichtigsten Informationen zu erkennen.	During this task, it was exhausting to find the important information.
ECL	Die Darstellung bei dieser Aufgabe ist ungünstig, um wirklich etwas zu lernen.	The design of this task was very inconvenient for learning.
ECL	Bei dieser Aufgabe ist es schwer, die zentralen Inhalte miteinander in Verbindung zu bringen.	During this task, it was difficult to recognize and link the crucial information.

*ICL, intrinsic cognitive load; ECL, extraneous cognitive load; GCL, germane cognitive load. *Item only useful if GCL is varied on purpose in a given learning material (e.g., through prompts).*

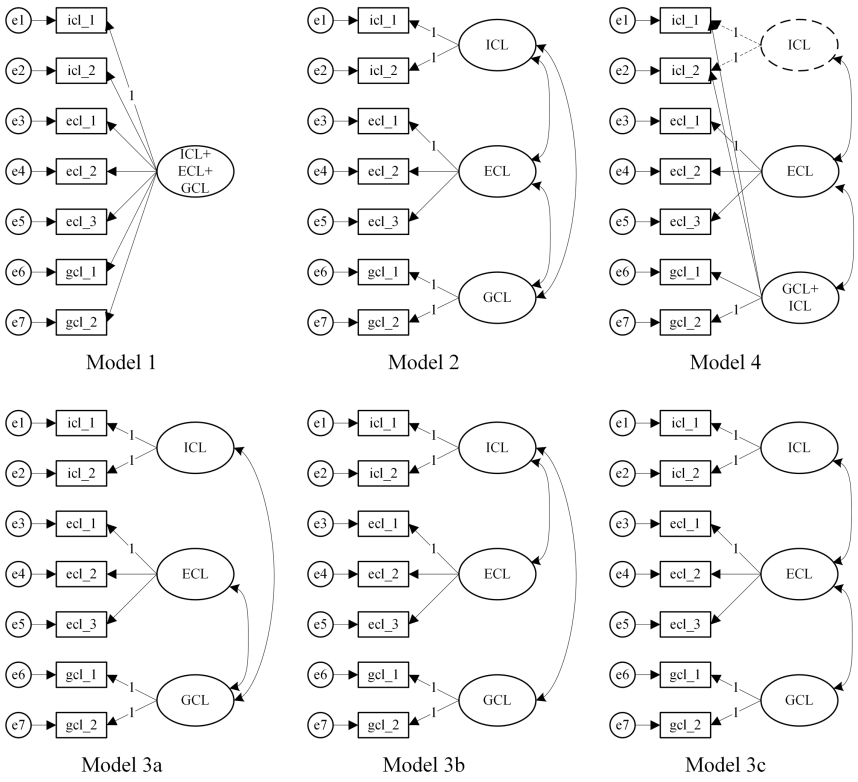

FIGURE 2 | Models for running the simultaneous scenario based factor analysis.

would be a worked example for the preceding ones. For solving all tasks, participants needed about 45 min. Unfortunately, some participants did not conduct all tasks (65 out of 95 participants completed all tasks).

Data Analysis

Again, we checked *reliability* in terms of *internal consistency* per task for each type of cognitive load and reported this data in an aggregated way. The germane-load scale was analyzed with a three-item version and a two-item version, where the item for instructional support to enhance germane load within the learning task was taken out, as corresponding instructional means are not always inherent in each learning environment. If Cronbach's alpha for the three-item GCL scale is not sufficient, all following analysis will be conducted with the two-item scale for GCL. *Validity* was analyzed like in Study 1 by testing whether the rating of the learners reflected the theoretical assumptions for each task. A confirmatory factor analysis simultaneously conducted across all tasks was used to test the *structure of the developed questionnaire.* The models, which have been tested, are shown in **Figure 2**. Model 1 assumes one unitary factor accounts for all load items. Model 2 represents the theoretical assumptions of three inter-related types of cognitive load and, therefore, suggests a three-factor model with separable but related cognitive load factors. Model 3a to 3c test if any of the latent relations between the three factors can be constrained to zero. Model 4 represents the revised perspective on cognitive load discussed by Kalyuga (2011) or Sweller et al. (2011): First, there is a productive load, comprising aspects of intrinsic and germane load, and second, there is an extraneous or unproductive load. Model 4 corresponds with this theoretical perspective: The ECL factor accounts for variance in ECL items. A broad GCL+ICL factor accounts for variance in both, GCL and ICL items. A nested ICL factor accounts for the ICL-specific variance in ICL items (i.e., technically, an ICL method specific factor). Adequate model fit is indicated by a low chi-square (χ^2) value, a high Tucker-Lewis index ($TLI \geq 0.95$), a high comparative fit index ($CFI \geq 0.95$), and a low root-mean-square error of approximation ($RMSEA \leq 0.05$). Additionally, the Akaike information criterion (*AIC*) was used as a model comparison index: The model yielding the lowest *AIC* is preferred in terms of close model fit and parsimony of the model relative to competing models.

Results

To analyze *reliability* of the three cognitive load subscales we calculated the *internal consistency* for each task. Reliability generalization showed that the mean Cronbach's α was 0.81 for the ICL scale, $\alpha = 0.86$ for the ECL scale, and $\alpha = 0.67$ for the GCL scale with three items. If the item that asked for instructional elements that support comprehension has been excluded, the remaining two-item scale for GCL resulted in an aggregated α of 0.85. If we only look at tasks that induce GCL through special elements in the learning material (e.g., prompts) the three-item version revealed an aggregated α of 0.70, as a result of more variance in the items for measuring GCL. Therefore, all following analysis were conducted using the two-item scale for GCL.

Validity was analyzed by comparing the ratings of the learners with the theoretically predicted outcomes, as we designed all 17 tasks to be related to high or low ICL, ECL, and GCL. Replicating results obtained in study 1, we found significantly different ratings for the low versus high load groups of tasks for each type of load [ICL: $t(76) = 7.85$, $p < 0.001$, $d = 0.73$; ECL: $t(89) = 7.28$, $p < 0.001$, $d = 0.94$; GCL: $t(76) = 3.39$, $p < 0.01$, $d = 0.35$]. Thereby, all scales were able to differentiate as expected (see **Figure 3** and **Table 4**).

As a next step, to get a closer look at the *structure of the developed questionnaire*, we conducted a confirmatory factor analysis with multiple groups. Model fits of all models can be found in **Table 5**. Overall, Model 2, representing the traditional view of three types of cognitive load, had the best fit. Conversely, Model 1 revealed the worst fit. Model 3a to 3c were inferior to Model 2. Model 4, with the recently discussed only two types of load, does reveal a decent fit. However, due to model complexity, i.e., the number of parameters to be estimated, the model comparison index AIC suggested Model 2 having the preferred fit. In conclusion, Model 2 was found to offer the best solution, thereby supporting three inter-related factors of cognitive load.

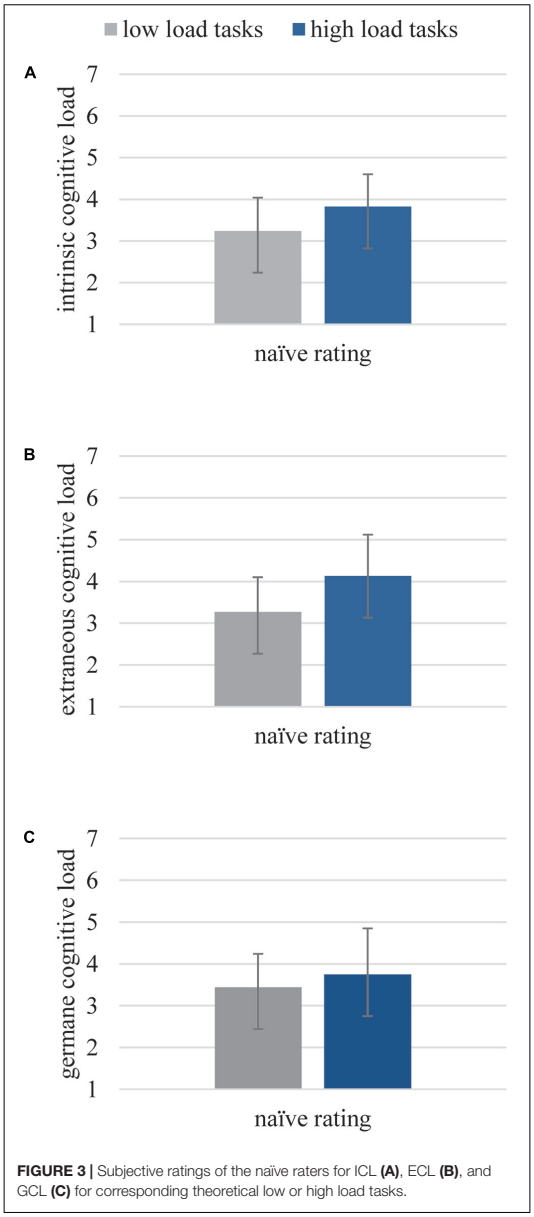

FIGURE 3 | Subjective ratings of the naïve raters for ICL **(A)**, ECL **(B)**, and GCL **(C)** for corresponding theoretical low or high load tasks.

Discussion

Study 2 showed that the naïve rating with our questionnaire seems to be a promising way of assessing cognitive load. The two scales on ICL and ECL again showed satisfying reliability and validity scores, as previously shown in Study 1. After substitution of the GCL items, this scale also appeared to be sensitive as intended. Cronbach's alpha was only moderate for the three-item version on germane load, when including one item for implemented instructional means to support learners' activities. Internal consistency could be improved if this item were removed. However, this specific item ("The learning task consisted of elements supporting my comprehension of the task.") is appealing because of its face validity for learning material including worked examples or prompts. Leppink et al. (2014) discuss a similar problem, as their original four items focus on the influence of the carried-out activity. Adding one item, for invested mental effort considerably decreased internal consistency. Overall, the questionnaire had considerably good item and scale characteristics. The sensitivity of the scales to detect the theoretically implemented load variations in various testing tasks especially leads us to the conclusion that the questionnaire can be broadly used in multimedia learning research.

Concerning the question of how many load types should be captured, our confirmatory factor analysis suggested

TABLE 4 | Means (M) and standard deviations (SD) for subjective ratings of the naïve raters for ICL, ECL, and GCL for corresponding theoretical low or high load tasks.

Type of load		Naïve raters	
		Low load tasks	High load tasks
ICL	*M (SD)*	3.24 (0.80)	3.82 (0.78)
ECL	*M (SD)*	3.27 (0.83)	4.13 (0.99)
GCL	*M (SD)*	3.44 (0.80)	3.75 (1.01)

ICL, intrinsic cognitive load; ECL, extraneous cognitive load; GCL, germane cognitive load.

that the traditional view of three interrelated types of cognitive load factors offers the best fit for our self-report questionnaire. However, a competing model based on the current view on CLT with only two factors (and an additional item-specific method factor) followed closely, and revealed decent fit as well. Nevertheless, we prefer GCL to be considered as a separable factor because the scale was shown to reflect variation of a germane-load variation in the generated cognitive tasks. These items are especially important whenever learners are activated on purpose by instructional means. As many studies aim at analyzing such activating instructions, like using prompts or desirable difficulties, it seems of high value to measure whether learners really follow these instructions and actually engage in the learning process.

GENERAL DISCUSSION

The goal of our two studies was to develop a reliable, valid, and practicable instrument to measure cognitive load in a differentiated way. The current extensively used method to measure cognitive load with only one item asking for "invested mental effort" (Paas, 1992) is from a methodological view not sufficient. Furthermore, the resulting problem of not knowing which aspect of load really was enhanced and afterward inferring the possible source of load in relation to learning outcomes is in our view also not satisfying.

Benefits of Measuring Load Differentially

With the developed questionnaire, we try to overcome the problem of inferring the source of cognitive load by using a questionnaire with several items that directly measures the constitutional parts of cognitive load, i.e., ICL, ECL, and GCL.

The differentiation can first be of interest to better understand individual learning processes. When learners with different prerequisites, such as prior knowledge, memory capacity, learning strategies, etc., deal with the same learning task, this might lead to different levels of different types of load during learning. Following, a differential approach might elect the sources of cognitive load and in addition predict learning outcome via regression analysis. Especially for comparisons of experts and novices, it can be fruitful to better understand their specific use of resources and their perception of loading sources. Furthermore, expertise reversal effects could be more easily explained and ascribed to specific types of load.

Second, the differentiated measurement can also be of great use to better understand the effects of instructional means that are not yet fully understood theoretically, like the effects of desirable difficulties. From a cognitive load point of view, one might expect that difficulties should lead to worse learning outcomes, but they actually can even enhance learning. On which type of load this fostering effect can be attributed is not yet clear but could be answered when using a differentiated load questionnaire.

We do not want to engage in the theoretical discussion of how many load types we should consider from a theoretical point of view. Instead, we argue that for *measuring* cognitive load, it can be fruitful to measure ICL, ECL, and GCL differentiated, even if the nature of germane load was questioned (Schnotz and Kürschner, 2007; de Jong, 2010; Moreno, 2010). Therefore, Ayres (2011), Kalyuga (2011), and Sweller et al. (2011) suggest rethinking the concept of GCL: It is argued that GCL, other than ICL or ECL, is not imposed by the learning material. Rather, they think that there are germane resources needed in working memory that need to be allocated to deal with the intrinsic load resulting from the learning material. There is no statement made by Sweller et al. (2011), if it is possible to allocate much more germane resources than necessary to deal with the materials' ICL. These allocated germane resources can be used for schema acquisition and deeper understanding and additionally for elaboration and connections to prior knowledge. When thinking to an end, they are nothing other than GCL according to the traditional view (Moreno and Park, 2010) of load. However, one likes to name them, we end up measuring three types of cognitive load. In fact, this is our approach: we try to measure all

TABLE 5 | Fit indices for competing structural models of cognitive load.

	$\chi 2$	*df*	*p*	*TLI*	*CFI*	*RMSEA*	*AIC*
Model 1	1709.152	238	<0.001	0.418	0.612	0.071	2,185.152
Model 2	335.106	221	<0.001	0.951	0.970	0.021	845.106
Model 3a	613.961	238	<0.001	0.851	0.901	0.036	1,089.961
Model 3b	406.207	238	<0.001	0.933	0.956	0.024	882.207
Model 3c	593.886	238	<0.001	0.859	0.906	0.035	1,069.886
Model 4	283.891	187	<0.001	0.951	0.974	0.021	861.891

processes that could require working memory resources and lead to cognitive load in some way as they are theoretically modeled in papers on CLT (e.g., Moreno and Park, 2010; Sweller, 2010a). This means that load resulting from the complexity of the learning material (especially element interactivity), load managed though prior knowledge, load as an effect of the instructional design, which could either be unproductive (e.g., split attention) or productive (e.g., prompts), and resources invested by the learner resulting in load. For all these aspects, items were created and analyzed with respect to their individual factor structure. The evaluated models in Study 2 revealed that the model with three types of load (Model 2) and the model with two types of load with the germane-load items as a part of the intrinsic-load scale (Model 4) do not differ that much, only with Model 2 resulting in minimal better model fit. Hence, from a theoretical point of view, the results provide evidence for both approaches, from a measurement point of view the three-partite model seems to be more adequate, especially when considering the reliability and validity results.

Strengths and Weaknesses of Our Measures

Overall, the results of the presented studies imply that it is possible to measure cognitive load reliably and valid in a differentiated way. The informed rating from Study 1 especially seems to be a promising instrument to assess the different aspects of cognitive load. The downside of the informed rating strategy is that an introduction into CLT needs to be given to learners and study participants beforehand. This might result in a loss of test efficiency in a situation where the research objective is focused on testing the functionality of, e.g., a training on learning strategies. However, the method of informed rating might be an adequate instrument for analyzing cognitive aspects of instructional design of learning material and provide a promising alternative method to access the different load types occurring during working with a specific learning material. Additionally, a combination of informed and naïve rating would be a very interesting and promising way. The informed rating might be used in an early stage of learning material development. The ratings of informed experts or semi-experts, who got a standardized introduction, approve that the material is varying the intended type of load. As a next step, the naïve rating can be applied to assess the learner's actual cognitive load during the learning process.

The naïve rating's advantage is that it is easier to utilize and does not need as much time and cognitive investment, as no introduction to cognitive load is needed. The developed items are easy to understand and fast to answer and can be used in a variety of research projects in the field of instructional design and multimedia learning. As already mentioned, an adaption to the situation can be useful. From a methodological point of view, we must state that we did not test a huge variety of items, as is common when designing a questionnaire. This results in our top-down approach in developing our items, as mentioned above. All items reflect theoretical aspects of the different types of load. Our items were developed and carefully worded based on given definitions and descriptions of types of load through various researchers (e.g., Moreno and Park, 2010; Sweller, 2010a). Based on the results of Study 1, we state that the concepts of intrinsic and extraneous load are easy to operationalize through different items. However, Study 1 also showed us that this does not apply to the concept of GCL, mainly because of the different possibility of understanding the items. Therefore, the pilot study for Study 2 used a bottom-up approach in a first step: nine items were generated from experts. The aim was to cover various aspects of germane processes and find items that could be easily understood by learners. As a second step, we then extracted the three best-fitting items that operationalize influences on GCL by analyzing reliability and whether they differentiated well between tasks that were theoretically meant to induce either low or high levels of GCL. At this point, our recommendation for the GCL item "The learning task consisted of elements supporting my comprehension of the task" would be that it should be included based on the learning material used (whether there is instructional support for investing germane resources or not) and a test on reliability of the items on germane load to eventually exclude this item at the end of your study. We expect reliability to be good if a variation of GCL is applied on purpose in at least one group of learners in an experiment. Otherwise, reliability might be low, but it should be good if the item is excluded. Further analysis on this point is in progress.

With the newly developed differentiated questionnaire, we have to ask ourselves about the added value with respect to other existing differentiated scales, like the one from Leppink et al. (2013, 2014). Their differentiated questionnaire, in contrast to ours, is especially useful if you can clearly detect the most relevant concepts within your learning tasks, e.g., the main terms in a statistics course or a geography text. In their approach, those concepts are needed for rephrasing their items to fit the material and to analyze the load resulting from understanding these concepts. Our questionnaire is not that specific to the learning content and, therefore, only needs adoption based on the material you use, e.g., text, video, or podcast. For instance, if participants watch a learning video, it may be more appropriate to speak of "During watching this video, it was exhausting to find the important information" (instead of "During this task, it was exhausting to find the important information"). Therefore, the questionnaire is easy to apply and fits to each content, especially for short interventions like they are, e.g., analyzed in Mayers widely used multimedia learning materials (e.g., Mayer and Anderson, 1991, 1992; Mayer, 1997). Leppink et al. (2013) by themselves state that they developed the questionnaire to be used within complex knowledge domains. We developed our questionnaire to fit a wide variety of domains and studies. For longer interventions, complex learning materials, or longer learning times, we recommend using our questionnaire multiple times. This might also overcome the problem of overestimation of load as already mentioned and discussed by van Gog et al. (2011). We suggest using the questionnaire after well-selected points of time during the learning process, as it might be more meaningful when it is related to distinct parts of the learning material instead of using it after, e.g., every 10 min. We consider our questionnaire to be short enough to not interrupt learning unnecessarily.

The general problem remains that cognitive-load ratings are nevertheless self-reported measures. Learners need to be aware of their current state of cognitive resources and task demands. Regarding this, it could be interesting to add ratings of metacognitive skills to better understand learners' abilities to self-evaluate themselves. Such skills would comprise metacognitive knowledge of one's own memory system or of task demands and in addition the ability to monitor the learning progress during learning (Flavell, 1979). Based on their observations of the learning process learners who are skillful self-regulators might then also decide to adopt their learning behavior (Zimmerman, 1989). This naturally could influence learners perceived cognitive load. Consequently, a better understanding of learners' self-regulatory skills could enlighten the interplay between learners' dynamic experiences of load and their self-regulatory activities during learning as was recently discussed in a special issue on cognitive load and self-regulation (for an overview see de Bruin and Van Merriënboer, 2017).

First Evidence of Validity and Future Directions

From a methodological point of view, we must state, that the tasks we used to impose the different types of load for validation have been of short duration and were embedded in an artificial learning context. Therefore, we started to evaluate the questionnaire with a study program where we explicitly varied ICL, ECL, and GCL in real learning settings with greater samples. This approach additionally has the advantage that learning outcomes can be derived. Overall, with these studies we aim to prove the validity of our instrument. Again, for the design of these studies, we used a highly systematic approach with varying domains in which only one type of load is addressed to be varied on purpose. ICL should be varied especially in terms of element interactivity, whereas ECL variations should be covering a wide variety of multimedia principles. For GCL, a variation of instructional help through, e.g., prompts, should be considered. Also, variations covering motivation and personal involvement seem to be of value to better understand the interplay of motivational or affective states and cognitive load. First results are promising in terms of a good sensitivity of our instrument regarding the intended load variations—in type and level—as the two following studies demonstrate.

A study of Seufert et al. (2016) already used our questionnaire in an experiment on increasing disfluency levels. Their main result was that slightly to medium levels of disfluency can foster learning compared to a fluent text, but that learning outcome gets worse when disfluency gets too intense. With respect to cognitive load, they found no effects on intrinsic load as expected. However, extraneous load was on a medium level for all disfluency groups, despite the group with the highest level of disfluency, which has been rated as highly loading extraneously. Germane load, which was called engagement in their study, on the other hand, increases steadily with increasing disfluency and remains high even on the highest disfluency level. Nevertheless, this engagement on the highest disfluency level does not pay off, as extraneous load is also too high. This is especially interesting as germane load—against theoretical assumptions—in this case, is related to low learning outcomes. Only when taking all aspects of load into account, can one understand the whole picture. Leppink et al. (2014) decided to drop their scale on GCL because they did not find an acceptable correlation between learning outcome and invested germane load. In our view, this positive relation is not *per se* to be expected as the interplay between different load types; the overall amount of load and especially the individual skills affect whether enhanced germane processes really are successful.

Another study using the presented naïve questionnaire was conducted by Rogers et al. (2014). They investigated the different systems for learning piano. The developed system, P.I.A.N.O., with its novel roll notation directly onto the piano keys, should avoid split attention and motivate learners through fast learning success. It resulted in the most accurate performance while playing a piece of music compared to a group using a standard sheet notation and a group using the software Synthesia to learn the piece of music. The projected roll notation also reduced perceived intrinsic load significantly and the therefore avoided split attention also reduced extraneous load. The novel roll notation also resulted in significant higher germane load, than in the group with the standard sheet notation. In a user experience test, they also found P.I.A.N.O. to be ranked highest among their tested systems. Based on GCL ratings and the measured user experience, Rogers et al. (2014) assume a motivational factor to be relevant. Motivation is also stated as important to learning and cognitive load by Zander (2010). Zander shows that motivational prerequisites of learners influence cognitive load: If a learner is motivated, allocation of resources in working memory to deal with ICL, ECL, and GCL is appropriate for task difficulty. If a learner is not motivated, fewer resources are allocated to deal with cognitive load. If resources are limited already through insufficient motivation, ICL and ECL might block all available resources, so nothing more is left for GCL. Unfortunately, Zander tried, but wasn't successful, in measuring cognitive load in a differentiated way. A replication of her study with our naïve rating would be a great possibility to get more insights into the relationship between cognitive load and the motivational prerequisites of learners.

What's next? For further analyses and improvement, the naïve rating questionnaire will be implemented and tested in different learning situations and with a variation of multimedia effects and domains as mentioned above. This should provide further evidence that our questionnaire is sensitive to different variations of different load types and, therefore, valid. With such a reliable, valid, and easy-to-use measure, which is used in many different studies with many different learning tasks and learner types, we should be able to learn more about the nature of cognitive load during learning and hence to theoretically improve CLT in the long run.

ETHICS STATEMENT

This study was exempt from an ethic committee approval due to the recommendations of the German Research Association: All

subjects were in no risk out of physical or emotional pressure, we fully informed all subjects about the goals and process of this study and none of the subjects were patients, minors or persons with disabilities. In all studies participation was voluntary and all subjects signed a written informed consent and were aware that they had the chance to withdraw their data at any point of the study.

AUTHOR CONTRIBUTIONS

MK and TS contributed to the conception and design of the studies. MK developed the used material and questionnaires and led data collection for all studies. TS revised the questionnaires. MK and FS analyzed and interpreted the data. MK drafted the work, which was revised critically by FS and TS. All authors provided approval of the final submitted version of the manuscript and agree to be accountable for all aspects of the work in ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpsyg.2017.01997/full#supplementary-material

REFERENCES

Antonenko, P. D., Paas, F., Grabner, R., and van Gog, T. (2010). Using electroencephalography to measure cognitive load. *Educ. Psychol. Rev.* 22, 425–438. doi: 10.1007/s10648-010-9130-y

Ayres, P. (2006). Using subjective measures to detect variations of intrinsic cognitive load within problems. *Learn. Instr.* 16, 389–400. doi: 10.1016/j.learninstruc.2006.09.001

Ayres, P. (2011). "Rethinking germane cognitive load," in *Proceedings of the EARLI Conference 2011 "Education for a Global Networked Society": Book of Abstracts and Extended Summaries*, ed. European Association for Research on Learning and Instruction (Exeter: University of Exeter), 1768–1770.

Ayres, P., and Sweller, J. (2005). "The split-attention principle in multimedia learning," in *The Cambridge Handbook of Multimedia Learning*, ed. R. E. Mayer (Cambridge, NY: Cambridge University Press), 135–146.

Baddeley, A. D. (2003). Working memory: looking back and looking forward. *Nat. Rev. Neurosci.* 4, 829–839. doi: 10.1038/nrn1201

Berthold, K., Eysink, T. H. S., and Renkl, A. (2009). Assisting self-explanation prompts are more effective than open prompts when learning with multiple representations. *Instr. Sci.* 37, 345–363. doi: 10.1007/s11251-008-9051-z

Berthold, K., Nückles, M., and Renkl, A. (2007). Do learning protocols support learning strategies and outcomes? The role of cognitive and metacognitive prompts. *Learn. Instr.* 17, 564–577. doi: 10.1016/j.learninstruc.2007.09.007

Bratfisch, O., Borg, G., and Dornic, S. (1972). *Perceived Item-difficulty in Three Tests of Intellectual Performance Capacity*. Report No. 29. Stockholm: Institute of Applied Psychology.

Brünken, R., Plass, J. L., and Leutner, D. (2003). Direct measurement of cognitive load in multimedia learning. *Educ. Psychol.* 38, 53–61. doi: 10.1207/S15326985EP3801_7

Brünken, R., Plass, J. L., and Leutner, D. (2004). Assessment of cognitive load in multimedia learning with dual-task methodology: auditory load and modality effects. *Instr. Sci.* 32, 115–132.

Brünken, R., Seufert, T., and Paas, F. G. W. C. (2010). "Measuring cognitive load," in *Cognitive Load Theory*, eds J. L. Plass, R. Moreno, and R. Brünken (Cambridge: Cambridge University Press), 181–202.

Brünken, R., Steinbacher, S., Plass, J. L., and Leutner, D. (2002). Assessment of cognitive load in multimedia learning using dual-task methodology. *Exp. Psychol.* 49, 109–119. doi: 10.1027//1618-3169.49.2.109

Chandler, P., and Sweller, J. (1996). Cognitive load while learning to use a computer program. *Appl. Cogn. Psychol.* 10, 151–170. doi: 10.1002/(SICI)1099-0720(199604)10:2<151::AID-ACP380>3.0.CO;2-U

Cierniak, G. (2011). *Facilitating and Inhibiting Learning by the Spatial Contiguity of Text and Graphic: How Does Cognitive Load Mediate the Split-Attention and Expertise Reversal Effect?* Ph.D. dissertation, Eberhard Karls Universität Tübingen, Tübingen.

Cierniak, G., Scheiter, K., and Gerjets, P. (2009). Explaining the split-attention effect: is the reduction of extraneous cognitive load accompanied by an increase in germane cognitive load? *Comput. Hum. Behav.* 25, 315–324. doi: 10.1016/j.chb.2008.12.020

de Bruin, A. B. H., and van Merriënboer, J. J. G. (2017). Bridging cognitive load and self-regulated learning research: a complementary approach to contemporary issues in educational research. *Learn. Instr.* 51, 1–9. doi: 10.1016/j.learninstruc.2017.06.001

de Jong, T. (2010). Cognitive load theory, educational research, and instructional design: some food for thought. *Instr. Sci.* 38, 105–134. doi: 10.1007/s11251-009-9110-0

DeLeeuw, K. E., and Mayer, R. E. (2008). A comparison of three measures of cognitive load: evidence for separable measures of intrinsic, extraneous, and germane load. *J. Educ. Psychol.* 100, 223–234. doi: 10.1037/0022-0663.100.1.223

Eysink, T. H. S., de Jong, T., Berthold, K., Kolloffel, B., Opfermann, M., and Wouters, P. (2009). Learner performance in multimedia learning arrangements. an analysis across instructional approaches. *Am. Educ. Res. J.* 46, 1107–1149. doi: 10.3102/0002831209340235

Flavell, J. H. (1979). Metacognition and cognitive monitoring: a new area of cognitive–developmental inquiry. *Am. Psychol.* 34, 906–911. doi: 10.1037/0003-066X.34.10.906

Fletcher, J. D., and Tobias, S. (2005). "The multimedia principle," in *The Cambridge Handbook of Multimedia Learning*, ed. R. E. Mayer (Cambridge: Cambridge University Press), 117–134.

Gerjets, P., Scheiter, K., and Catrambone, R. (2004). Designing instructional examples to reduce intrinsic cognitive load: molar versus modular presentation of solution procedures. *Instr. Sci.* 32, 33–58. doi: 10.1023/B:TRUC.0000021809.10236.71

Gerjets, P., Scheiter, K., and Catrambone, R. (2006). Can learning from molar and modular worked examples be enhanced by providing instructional explanations and prompting self-explanations? *Learn. Instr.* 16, 104–121. doi: 10.1016/j.learninstruc.2006.02.007

Goldstein, R., Bauer, L. O., and Stern, J. A. (1992). Effect of task difficulty and interstimulus interval on blink parameters. *Int. J. Psychophysiol.* 13, 111–117. doi: 10.1016/0167-8760(92)90050-L

Hart, S. G., and Staveland, L. E. (1988). "Development of NASA-TLX (task load index): results of empirical and theoretical research," in *Human Mental Workload*, eds N. Meshkati and P. A. Hancock (Amsterdam: Elsevier), 139–183.

Kalyuga, S. (2011). Cognitive load theory: how many types of load does it really need? *Educ. Psychol. Rev.* 23, 1–19. doi: 10.1007/s10648-010-9150-7

Kalyuga, S., Chandler, P., and Sweller, J. (1998). Levels of expertise and instructional design. *Hum. Factors* 40, 1–17. doi: 10.1518/001872098779480587

Kirschner, P. A., Ayres, P., and Chandler, P. (2011). Contemporary cognitive load theory research: the good, the bad and the ugly. *Comput. Hum. Behav.* 27, 99–105. doi: 10.1016/j.chb.2010.06.025

Leppink, J., Paas, F. G. W. C., van der Vleuten, C. P. M., van Gog, T., and van Merriënboer, J. J. G. (2013). Development of an instrument for measuring different types of cognitive load. *Behav. Res. Methods* 45, 1058–1072. doi: 10.3758/s13428-013-0334-1

Leppink, J., Paas, F. G. W. C., van Gog, T., van der Vleuten, C. P. M., and van Merriënboer, J. J. G. (2014). Effects of pairs of problems and examples on task performance and different types of cognitive load. *Learn. Instr.* 30, 32–42. doi: 10.1016/j.learninstruc.2013.12.001

Low, R., and Sweller, J. (2005). "The modality principle in multimedia learning," in *The Cambridge Handbook of Multimedia Learning*, ed. R. E. Mayer (Cambridge, NY: Cambridge University Press), 147–158.

Mayer, R. E. (1997). Multimedia learning. Are we asking the right questions?. *Educ. Psychol.* 32, 1–19. doi: 10.1207/s15326985ep3201_1

Mayer, R. E. (2001). *Multimedia Learning*. Cambridge, NY: Cambridge University Press.

Mayer, R. E. (2005a). "Principles for managing essential processing in multimedia learning: segmenting, pretraining, and modality principles," in *The Cambridge Handbook of Multimedia Learning*, ed. R. E. Mayer (Cambridge, NY: Cambridge University Press), 169–182.

Mayer, R. E. (ed.) (2005b). *The Cambridge Handbook of Multimedia Learning*. Cambridge: Cambridge University Press.

Mayer, R. E., and Anderson, R. B. (1991). Animations need narrations: an experimental test of a dual-coding hypothesis. *J. Educ. Psychol.* 83, 484–490. doi: 10.1037/0022-0663.83.4.484

Mayer, R. E., and Anderson, R. B. (1992). The instructive animation: helping students build connections between words and pictures in multimedia learning. *J. Educ. Psychol.* 84, 444–452. doi: 10.1037/0022-0663.84.4.444

Mayer, R. E., and Moreno, R. (2002). Aids to computer-based multimedia learning. *Learn. Instr.* 12, 107–119. doi: 10.1016/S0959-4752(01)00018-4

Mayer, R. E., and Moreno, R. (2003). Nine ways to reduce cognitive load in multimedia learning. *Educ. Psychol.* 38, 43–52. doi: 10.1207/S15326985EP3801_6

Mayer, R. E., and Moreno, R. (2010). "Techniques that reduce extraneous cognitive load and manage intrinsic cognitive load during multimedia learning," in *Cognitive Load Theory*, eds J. L. Plass, R. Moreno, and R. Brünken (Cambridge, NY: Cambridge University Press), 131–152.

Moreno, R. (2010). Cognitive load theory: more food for thought. *Instr. Sci.* 38, 135–141. doi: 10.1007/s11251-009-9122-9

Moreno, R., and Park, B. (2010). "Cognitive load theory: historical development and relation to other theories," in *Cognitive Load Theory*, eds J. L. Plass, R. Moreno, and R. Brünken (Cambridge, NY: Cambridge University Press), 9–28.

Naismith, L. M., Cheung, J. J. H., Ringsted, C., and Cavalcanti, R. B. (2015). Limitations of subjective cognitive load measures in simulation-based procedural training. *Med. Educ.* 49, 805–814. doi: 10.1111/medu.12732

Paas, F., Renkl, A., and Sweller, J. (2003a). Cognitive load theory and instructional design: recent developments. *Educ. Psychol.* 38, 1–4. doi: 10.1207/S15326985EP3801_1

Paas, F., Tuovinen, J. E., Tabbers, H., and van Gerven, P. W. M. (2003b). Cognitive load measurement as a means to advance cognitive load theory. *Educ. Psychol.* 38, 63–71. doi: 10.1207/S15326985EP3801_8

Paas, F. G. W. C. (1992). Training strategies for attaining transfer of problem-solving skill in statistics: a cognitive-load approach. *J. Educ. Psychol.* 84, 429–434. doi: 10.1037/0022-0663.84.4.429

Paas, F. G. W. C., and van Gog, T. (2006). Optimising worked example instruction: different ways to increase germane cognitive load. *Learn. Instr.* 16, 87–91. doi: 10.1016/j.learninstruc.2006.02.004

Paas, F. G. W. C., van Gog, T., and Sweller, J. (2010). Cognitive load theory: new conceptualizations, specifications, and integrated research perspectives. *Educ. Psychol. Rev.* 22, 115–121. doi: 10.1007/s10648-010-9133-8

Paas, F. G. W. C., and van Merriënboer, J. J. G. (1994). Variability of worked examples and transfer of geometrical problem-solving skills: a cognitive-load approach. *J. Educ. Psychol.* 86, 122–133. doi: 10.1037/0022-0663.86.1.122

Park, B., and Brünken, R. (2011). "The rhythm task - a new method for measuring cognitive load," in *Proceedings of the EARLI Conference 2011 "Education for a Global Networked Society": Book of Abstracts and Extended Summaries*, ed. European Association for Research on Learning and Instruction (Exeter: University of Exeter), 1059–1060.

Rey, G. D. (2012). A review of research and a meta-analysis of the seductive detail effect. *Educ. Res. Rev.* 7, 216–237. doi: 10.1016/j.edurev.2012.05.003

Rodriguez, M. C., and Maeda, Y. (2006). Meta-analysis of coefficient alpha. *Psychol. Methods* 11, 306–322. doi: 10.1037/1082-989X.11.3.306

Rogers, K., Röhlig, A., Weing, M., Gugenheimer, J., Könings, B., Klepsch, M., et al. (2014). "P.I.A.N.O," in *ITS '14 Proceedings of the Ninth ACM International Conference on Interactive Tabletops and Surfaces*, eds R. Dachselt, N. Graham, K. Hornbæk, and M. Nacenta (New York, NY: ACM), 149–158.

Rumelhart, D. E. (1981). "Schemata: the building blocks of cognition," in *Comprehension and Teaching: Research Reviews*, ed. J. T. Guthrie (Newark, DE: International Reading Assn), 3–26.

Salomon, G. (1984). Television is "easy" and print is "tough": the differential investment of mental effort in learning as a function of perceptions and attributions. *J. Educ. Psychol.* 76, 647–658. doi: 10.1037/0022-0663.76.4.647

Schnotz, W., and Kürschner, C. (2007). A reconsideration of cognitive load theory. *Educ. Psychol. Rev.* 19, 469–508. doi: 10.1007/s10648-007-9053-4

Seufert, T., Wagner, F., and Westphal, J. (2016). The effects of different levels of disfluency on learning outcomes and cognitive load. *Instr. Sci.* 45, 221–238. doi: 10.1007/s11251-016-9387-8

Swaak, J., and de Jong, T. (2001). Learner vs. System control in using online support for simulation-based discovery learning. *Learn. Environ. Res.* 4, 217–241. doi: 10.1023/A:1014434804876

Sweller, J. (1994). Cognitive load theory, learning difficulty, and instructional design. *Learn. Instr.* 4, 295–312. doi: 10.1016/0959-4752(94)90003-5

Sweller, J. (2010a). "Cognitive load theory: recent theoretical advances," in *Cognitive Load Theory*, eds J. L. Plass, R. Moreno, and R. Brünken (Cambridge: Cambridge University Press), 29–47.

Sweller, J. (2010b). Element interactivity and intrinsic, extraneous, and germane cognitive load. *Educ. Psychol. Rev.* 22, 123–138. doi: 10.1007/s10648-010-9128-5

Sweller, J., Ayres, P., and Kalyuga, S. (2011). *Cognitive Load Theory*. New York, NY: Springer.

Sweller, J., and Chandler, P. (1991). Evidence for cognitive load theory. *Cogn. Instr.* 8, 351–362. doi: 10.1207/s1532690xci0804_5

Sweller, J., and Chandler, P. (1994). Why some material is difficult to learn. *Cogn. Instr.* 12, 185–233. doi: 10.1207/s1532690xci1203_1

Sweller, J., van Merriënboer, J. J. G., and Paas, F. G. W. C. (1998). Cognitive architecture and instructional design. *Educ. Psychol. Rev.* 10, 251–296. doi: 10.1023/A:1022193728205

van Gerven, P. W. M., Paas, F. G. W. C., van Merriënboer, J. J. G., and Schmidt, H. G. (2004). Memory load and the cognitive pupillary response in aging. *Psychophysiology* 41, 167–174. doi: 10.1111/j.1469-8986.2003.00148.x

van Gog, T., Kirschner, F., Kester, L., and Paas, F. (2011). "When to measure cognitive load with mental effort rating scales," in *Proceedings of the EARLI Conference 2011 "Education for a Global Networked Society": Book of Abstracts and Extended Summaries*, ed. European Association for Research on Learning and Instruction (Exeter: University of Exeter), 1056–1057.

van Gog, T., and Paas, F. G. W. C. (2008). Instructional efficiency: revisiting the original construct in educational research. *Educ. Psychol.* 43, 16–26. doi: 10.1080/00461520701756248

VanLehn, K., Jones, R. M., and Chi, M. T. (1992). A model of the self-explanation effect. *J. Learn. Sci.* 2, 1–59. doi: 10.1207/s15327809jls0201_1

Whelan, R. R. (2007). Neuroimaging of cognitive load in instructional multimedia. *Educ. Res. Rev.* 2, 1–12. doi: 10.1016/j.edurev.2006.11.001

Wilson, G. F., and Eggemeier, F. T. (1991). "Physiological measures of workload in multi-task environments," in *Multiple-Task Performance*, ed. D. L. Damos (Washington, DC: Taylor & Francis), 329–360.

Zander, S. (2010). *Motivationale Lernervoraussetzungen in der Cognitive Load Theory [Motivational Prerequisites of Learners in Cognitive Load Theory]*. Berlin: Logos-Verl.

Zimmerman, B. J. (1989). "Models of self-regulated learning and academic achievement," in *Self-Regulated Learning and Academic Achievement*, eds B. J. Zimmerman and D. H. Schunk (New York, NY: Springer), 1–25.

Zumbach, J., and Mohraz, M. (2008). Cognitive load in hypermedia reading comprehension. Influence of text type and linearity. *Comput. Hum. Behav.* 24, 875–887. doi: 10.1016/j.chb.2007.02.015

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Pichler, M. & Seufert, T. (2011). Two strategies to measure Cognitive Load. In European Association for Research on Learning and Instruction (Hg.), *EARLI Conference 2011 Ëducation for a Global Networked Society": Book of Abstracts and Extended Summaries* (S. 928–929). University of Exeter.

Abrufbar unter: https://www.earli.org/book-abstracts

European Association for Research on Learning and Instruction (EARLI) Conference 2011 – Accepted Submission

Two strategies to measure Cognitive Load

Melina Pichler
Ulm University, Institute of Psychology and Education
Albert-Einstein-Allee 47, 89081 Ulm, Germany
melina.pichler@uni-ulm.de

Tina Seufert
Ulm University, Institute of Psychology and Education
Albert-Einstein-Allee 47, 89081 Ulm, Germany
tina.seufert@uni-ulm.de

Abstract
Despite the extensive use of concepts from Cognitive Load Theory, the problem of how to assess cognitive load, especially the intrinsic, extraneous and germane load aspects remains still unsolved. We developed two strategies to measure cognitive load: (1) First, we trained a group of learners in differentiating the three concepts before rating (informed rating). (2) Second, we developed a questionnaire with different items for each type of cognitive load (differential rating). The question was how valid, reliable and how comparable the two instruments are. In an experimental study (n=51) with 24 tasks each varying in one aspect of cognitive load learners of both groups had to rate the three types of load and additionally the overall amount of load. To analyse the reliability of the two instruments we calculated the overall internal consistency and for the differential rating additionally the internal consistency per load type. The validity was analyzed by comparing the ratings of the learners with the theoretically expected outcomes. Both instruments turned out to be valid, except the germane load scale of the differential rating. The overall internal consistency of the informed rating was weak, indicating that may be better for differential than overall load ratings. The overall and specific reliability of the differential rating instrument was good. In many tasks we found differences between the two instruments due to more distinct values of the informed rating group. Overall, especially the informed rating seems to be a promising strategy to assess different aspects of cognitive load.

Keywords: Assessment methods, Learning theory, New Modes of Assessment

Introduction and Theoretical Background

One of the most powerful and most criticized research frameworks in educational research of the last decades is the Cognitive Load Theory (CLT; Chandler & Sweller, 1991). Powerful because the concept is extensively used for evaluating learning environments or interpreting empirical results. Criticized, because there is at least one main problem that is still not solved adequately: how to measure cognitive load (Moreno, 2010). Almost all studies on multimedia learning assess cognitive load by using one item for rating the perceived invested mental effort (Paas, 1992). Some other studies used objective techniques like dual task measures or physiological parameters. However, there is no appropriate instrument to measure the three conceptual parts of CLT, the intrinsic, extraneous and germane cognitive load. In the present study we investigated two

different strategies for assessing these different parts of cognitive load: (1) First, we trained a group of learners in differentiating the three concepts before rating (informed rating). (2) Second, we developed a questionnaire with different items for each type of cognitive load (differential rating). The question was whether the two instruments were reliable and valid and whether they lead to comparable ratings of cognitive load when the same tasks had to be evaluated.

Method

In the experimental study 51 students took part and where randomly assigned to one of the two experimental groups. The informed rating group got a short introduction into cognitive load, especially the three different types of cognitive load, whereas the differential rating group didn't get any instruction. Then every learner had to evaluate 24 tasks where the three types of cognitive load had been varied based on theoretical assumptions and empirical findings: (a) For intrinsic cognitive load (ICL) the element-interactivity of a task had been varied, e.g. "the day after tomorrow will be Saturday. Which day was yesterday" versus "Three days after yesterday was Friday. Which day will be five days before tomorrow?" (b) For extraneous load (ECL) we showed learning environments with an integrated format of text and picture versus a separated format etc. (c) For germane load (GCL) we asked to rate different instructional settings which should either induce germane load, like "every 20 minute a teacher gives you time to think of examples you can find for the topic" versus tasks without such an activation. We created pairs or groups of tasks with the same content type, only varying in the amount of ICL, ECL or GCL. Students had to rate the three types of load in a questionnaire. The informed rating group had to answer one item per load type, the differential rating group got 2 respective 3 items for each load type: For ICL we asked for example how complex the task was. For ECL learners should rate whether it was exhausting to find the relevant information. For GCL we asked learners to rate whether this task animates to think intensively about the given topic. In both questionnaires learners additionally had to rate the overall amount of perceived cognitive load on a 7-point-likert-scale from very low to very high mental effort, which was adapted from Paas (1992).

To analyse the reliability of the two instruments we calculated the overall internal consistency for each task. For the differential rating test with more than one item per load type we also analyzed the internal consistency per task for each type of cognitive load. The validity was analyzed by comparing the ratings of the learners with the theoretically expected outcomes, i.e. the theoretically defined oppositional tasks should be rated significantly different in the respective type of load.

Results

The *reliability* of the informed rating test was quite poor for the different tasks: except for two tasks, the internal consistency was not sufficient. The differential rating test on the other hand had good to excellent values of α overall from .62 to .93. For the different load types the internal consistency was also good (with the exception 2 tasks) with α values about .80.

The *validity* test for the informed rating test revealed significantly different ratings for

the oppositional groups of tasks (extraneous load: $F(1,26) = 110.94$, $p<.001$, $\eta^2=.82$, germane load: $F(1,21) = 23.51$, $p<.001$, $\eta^2=.54$, intrinsic load: $F(1,22) = 117.86$, $p<.001$, $\eta^2=.85$). The differential rating test also differentiated very good for differences in extraneous load ($F(1,25) = 54.50$, $p<.001$, $\eta^2=.69$) and intrinsic load ($F(1,23) = 67.01$, $p<.001$, $\eta^2=.75$), but not that good for differences in germane load ($F(1,23) = 0.07$, $p=.79$, $\eta^2=.00$).

Moreover we analyzed whether there are *differences* between the ratings of the two groups. For ECL the two instruments were comparable for 73% of the tasks, for ICL only for 70% and for GCL the ratings were comparable only for 42% of the tasks. However, in all those cases where there are differences between the two groups, the informed rating group rated the load type of the task more distinct, e.g. for a low germane load task the germane load rating is much lower than in the differential rating group.

Summary and Discussion

The informed rating turned out to be a valid measure for the three types of cognitive load. However, the overall reliability was not sufficient, indicating that it is much better to make differential ratings instead of calculating an overall amount of cognitive load. Nevertheless, other reliability measures like retests should be considered. The differential rating instrument also was valid for ECL and ICL, but for GCL the items should be revised. The great differences between the two questionnaires are due to the more distinct ratings of the informed rating group. Overall, especially the informed rating seems to be a promising strategy to assess different aspects of cognitive load.

References

Chandler, P., & Sweller, J. (1991). Cognitive Load Theory and the Format of Instruction. *Cognition and Instruction, 8*, 293-332.

Paas, F. (1992). Training strategies for attaining transfer of problem-solving skill in statistics: A cognitive-load approach. *Journal of Educational Psychology, 84*, 429–434.

Moreno, R. (2010). Cognitive load theory: more food for thought. *Instructional Science, 38*, 135–141.

Klepsch, M. & Seufert, T. (9. April 2012). *Subjective Differentiated Measurement of Cognitive Load*. 5th International Cognitive Load Theory Conference, Tallahassee (USA).

International Cognitive Load Theory Conference 2012
Accepted Submission

Subjective Differentiated Measurement of Cognitive Load

Melina Klepsch
Ulm University, Institute of Psychology and Education
Albert-Einstein-Allee 47, 89081 Ulm, Germany
melina.klepsch@uni-ulm.de

Tina Seufert
Ulm University, Institute of Psychology and Education
Albert-Einstein-Allee 47, 89081 Ulm, Germany
tina.seufert@uni-ulm.de

Despite the popularity and extensive use of Cognitive Load Theory (CLT; Chandler & Sweller, 1991) in educational research, the problem of how to assess cognitive load, especially the intrinsic, extraneous and germane load aspects still remains unsolved (Moreno, 2010). Many studies assess cognitive load by using one item for rating the invested mental effort (Paas, 1992). Others use objective techniques like dual task measures. We developed a subjective questionnaire with different items for each type of cognitive load (differential rating). The question was how valid and reliable this instrument is.

In our experimental study 95 students took part. During an online-study they had to solve 17 learning and problem solving tasks, each varying the three conceptual parts of CLT. The variation based on theoretical assumptions and empirical findings: (a) For intrinsic cognitive load (ICL) we varied the element-interactivity of a task. (b) For extraneous load (ECL) we presented for example learning environments with an integrated format of text and picture versus a separated format. (c) For germane load (GCL) we showed learning tasks which should either induce germane load by activating deeper learning processes versus tasks without such an activation. We clustered groups of tasks addressing the same type of load, only varying in the amount of load (theoretically low versus high).

After each task, every participant had to rate cognitive load. Our questionnaire consists of 2 items on ICL, 3 items on ECL and 3 items on GCL and is an improvement of a questionnaire developed a year ago (Pichler & Seufert, 2011). For ICL we asked for example how complex the task was. For ECL learners should rate whether it was exhausting to find the relevant information. For GCL we asked learners to rate whether they were engaged in understanding the whole task and not only parts of it. All items had to be answered on a 7-point-likert-scale (absolutely wrong to absolutely right).

To analyse the *reliability* of the three cognitive load subscales we calculated their internal consistency for every point of measurement, i.e. for each task. The analysis of internal consistency showed one critical item in the GCL scale, which has therefore been excluded. This resulted in good values of α for each subscale in nearly every task, as you can see in table 1.

Next, the *validity* was analysed by comparing the ratings of the learners with theoretically expected outcomes, as we designed all 17 tasks to be related to high or low ICL, ECL and/or GCL. This classification was validated by experts in a pre-test.

As expected we found significantly different ratings for the oppositional groups of tasks (intrinsic load: $t(76) = 7.85$, $p<.001$, $d=.73$, extraneous load: $t(89) = 7.28$, $p<.001$, $d=.94$, germane load: $t(76) = 3.39$, $p<.01$, $d=.35$).

The differential rating seems to be a promising way of assessing the three conceptual parts of CLT. Nevertheless we need to test our questionnaire in "real" learning and problem solving tasks with a greater sample of subjects and tasks and in varying domains.

References

Chandler, P., & Sweller, J. (1991). Cognitive Load Theory and the Format of Instruction. *Cognition and Instruction, 8*, 293-332.

Moreno, R. (2010). Cognitive load theory: more food for thought. *Instructional Science, 38*, 135–141.

Paas, F. (1992). Training strategies for attaining transfer of problem-solving skill in statistics: A cognitive-load approach. *Journal of Educational Psychology, 84,* 429–434.

Pichler, M., & Seufert, T. (2011). *Two strategies to measure Cognitive Load*. EARLI Conference 2011 "Education for a Global Networked Society" (30.08.-03.09.2011), Exeter.

Table 1. Cronbach's Alpha of intrinsic, extraneous and germane load scale for all 17 tasks.

Task	**Cronbach's Alpha**		
	Intrinsic Load (2 Items)	Extraneous Load (3 Items)	Germane Load (2 Items)
01	.744	.593	.621
02	.712	.811	.866
03	.816	.868	.628
04	.843	.720	.797
05	.924	.902	.759
06	.705	.881	.863
07	.837	.858	.624
08	.779	.863	.814
09	.761	.831	.785
10	.752	.845	.891
11	.831	.891	.829
12	.820	.825	.861
13	.705	.919	.954
14	.762	.842	.651
15	.806	.853	.818
16	.846	.877	.910
17	.634	.854	.712

4. Praktischer Einsatz des Fragebogens

Zur Validierung und um die Relevanz für Instruktionsdesignfragestellungen klar zu machen wurde der Fragebogen in unterschiedliche Settings durch eine Serie von sechs Studien getestet. Zwei Studien variieren die Komplexität des Lerninhaltes um eine Variation von ICL zu erzeugen, zwei Studien variieren die Darstellung des Inhaltes um eine Variation von ECL zu erzeugen und zwei Studien variieren die Aktivierung der kognitiven Prozesse des Lernenden um eine Variation von GCL zu erzeugen. Durch die Zusammenfassung der sechs Studien in einem Zeitschriftenbeitrag war es möglich ein umfassendes Bild der Relevanz einer differenzierten Messung der kognitiven Belastung beim Lernen für Instruktionsdesignfragestellungen darzustellen.

- Klepsch, M. & Seufert, T. (2020). Understanding instructional design effects by differentiated measurement of intrinsic, extraneous, and germane cognitive load. *Instructional Science, 48*(1), 45–77. doi: 10.1007/s11251-020-09502-9

Einzelne Studien aus der Serie waren zudem teilweise in einen größeren Kontext eingebettet und weitere Experimentalgruppen wurden erfasst. Außerdem wurde eine weitere Studie zur Variation von ICL durchgeführt. Diese Ergebnisse wurden auf Tagungen präsentiert und sind teilweise in Tagungsbänden erschienen. Um diesen zusätzlichen Analysen und Ergebnissen gerecht zu werden wurden daher folgende Beiträge integriert:

- Klepsch, M., Kempter, I. & Seufert, T. (2013). Interaction of Worked Examples and Prompts: Impact on Performance and Cognitive Load. In European Association for Research on Learning and Instruction (Hrsg.), *15th Biennial EARLI Conference "Responsible Teaching and Sustainable Learning"*, München (DE).

- Klepsch, M., Westphal, J. & Seufert, T. (2013). Effekte von integriertem bzw. separiertem Text-Bild-Material in Abhängigkeit von serialistischem oder holistischem Lernstil., *14. Fachgruppentagung Pädagogische Psychologie*. Hildesheim (DE).

- Seufert, T., Klepsch, M. & Westphal, J. (2017). Adjusting Task Difficulty to Control Learner's Intrinsic Cognitive Load. In European Association for Research on Learning and Instruction (Hrsg.), *EARLI 2017 Ëducation in the crossroads of economy and politics - Role of research in the advancement of public good"*, Tempere (FIN).

Eine Studie bei der alle Arten kognitiver Belastung variiert wurden ist in Kooperation mit dem Institut für Medieninformatik der Universität Ulm entstanden. Studierende der Medieninformatik entwickelten ein Programm zum Klavierspielen lernen ohne dabei auf das Notenlesen angewiesen zu sein. Die zwei Studien zur Evaluation der Software wurden auf einer Tagung vorgestellt und publiziert. Zudem wurden erweiterte Analysen mit nur zwei Experimentalgruppen mit Fokus auf den entwickelten Fragebogen zur differenzierten Messung kognitiver Belastung auf einer weiteren Tagung vorgestellt.

- Rogers, K., Röhlig, A., Weing, M., Gugenheimer, J., Könings, B., Klepsch, M., Schaub, F., Rukzio, E., Seufert, T. & Weber, M. (2014). P.I.A.N.O: Faster Piano Learning with Interactive Projection. In R. Dachselt, N. Graham, K. Hornbæk & M. Nacenta (Hrsg.), *Ninth ACM International Conference on Interactive Tabletops and Surfaces*, Dresden (DE). doi: 10.1145/2669485.2669514

- Klepsch, M., Könings, B., Weber, M. & Seufert, T. (25. August 2014). Fostering Piano Learning by dynamic mapping of notes. European Association for Research on Learning and Instruction SIG 2 (Hrsg.), *EARLI SIG 2 Meeting "Building bridges: Improving our understanding of learning from text and graphics by making the connection"*, Rotterdam (NE).

Klepsch, M. & Seufert, T. (2020). Understanding instructional design effects by differentiated measurement of intrinsic, extraneous, and germane cognitive load. *Instructional Science, 48*(1), 45–77. doi: 10.1007/s11251-020-09502-9

Abrufbar unter: https://doi.org/10.1007/s11251-020-09502-9

Instructional Science (2020) 48:45–77
https://doi.org/10.1007/s11251-020-09502-9

ORIGINAL RESEARCH

Understanding instructional design effects by differentiated measurement of intrinsic, extraneous, and germane cognitive load

Melina Klepsch[1] · **Tina Seufert**[1]

Received: 13 March 2019 / Accepted: 18 January 2020 / Published online: 4 February 2020

Abstract
Instructional design deals with the optimization of learning processes. To achieve this, three aspects need to be considered: (1) the learning task itself, (2) the design of the learning material, and (3) the activation of the learner's cognitive processes during learning. Based on Cognitive Load Theory, learners also need to deal with the task itself, the design of the material, and the decision on how much to invest into learning. To link these concepts, and to help instructional designers and teachers, cognitive load during learning needs to be differentially measured. This article reviews studies using a questionnaire to measure intrinsic, extraneous and germane cognitive load in order to provide evidence for the instruments' prognostic validity. Six exemplary studies from different domains with different variations of the learning material were chosen to show that the theoretically expected effects on different types of load are actually reflected in the learners' answers in the questionnaire. Major hypotheses regarding the different load types were (1) variations in difficulty are reflected in the scale on intrinsic cognitive load, (2) variations in design are reflected in the scale on extraneous cognitive load, and (3) variations in enhancing deeper learning through activation of cognitive processes are reflected in the scale on germane cognitive load. We found prognostic validity to be good. The review concludes by discussing the practical and theoretical implications, as well as pointing out the limitations and needs for further research.

Keywords Cognitive load theory · Differentiated measurement · Instructional design · Multimedia research · Multimedia design principles

✉ Melina Klepsch
melina.klepsch@uni-ulm.de

Tina Seufert
tina.seufert@uni-ulm.de

[1] Institute of Psychology and Education, Department Learning and Instruction, Ulm University, Albert-Einstein-Allee 47, Ulm 89081, Germany

Introduction and theoretical background

Instructional Design (ID) deals with the goal of optimizing learning processes. However, what is optimal mainly depends on the goals of learning. Based on the given goals, instructional design focuses on the question "What needs to be presented and how should it be presented to the learner to reach these goals". There are approaches that focus (1) on the learning task itself, (2) on the design of the learning material, and (3) on the activation of the learner's cognitive processes to invest effort into learning.

Approaches that focus on the learning task itself. Deciding on the learning content and the task itself is one of the most important parts instructors need to think about. Based on these decisions, instructors identify the prior knowledge needed to understand the content and the complexity of the task to be completed. Besides adjusting the task complexity, one could also enhance learners' prior knowledge by providing pre-training. The pre-training principle focuses on equipping "the learner with knowledge that will make it easier to process the lesson" (Mayer 2009, p. 190), which means familiarizing the learner with vocabulary and characteristics of key components of the learning content to follow. Another technique is reducing complexity, especially element interactivity, by presenting the material through an isolated elements procedure (Pollock et al. 2002; Ayres 2006), meaning that in a first step, complexity is reduced by de-constructing the task, and therefore learning is conducted element-by-element to allow learners to construct partial schemas, which can be combined in a second step. The goal of both described techniques is an increase of prior knowledge before the actual task needs to be performed.

Approaches that focus on the design of the learning material. Widely and systematically investigated are differences in the design of learning material and their effects on learning. There exists a large variety of principles and effects stating that one representation of learning material outperforms the other when it comes to measures of learning outcome or task performance. For an overview, see Mayer (2005b, 2009). All of these principles state dos and don'ts that should be considered when designing learning material. For example, adding a relevant picture to a text improves learning more than just presenting the text alone (multimedia principle; Fletcher and Tobias 2005), integrating the text into the picture would result in an even better learning outcome or performance than presenting both representations separately (split-attention effect; Kalyuga et al. 1998). Alternatively, the given text could be presented in an auditory format to avoid overloading working memory (modality principle; Moreno and Mayer 1999). In the majority of cases, the goal of these techniques is to avoid unnecessary search or navigation processes between different representations or overload when, for example, using just one modality.

Approaches that focus on the activation of the learner's cognitive processes to invest effort into learning. A range of approaches follow the idea of cognitive activation in order to "direct learners' attention to cognitive processes that are relevant for learning or schema construction" (Sweller et al. 1998, p. 262). These approaches often use cognitive or metacognitive prompts to hustle the learner in the right direction or to give the learner cues on how to best process the given learning content (prompting effect; Bannert 2009). The goal of these principles is to foster the construction of schemata and mental models.

Besides the question of whether learners perform better through one or more of these approaches, (e.g., achieve higher learning outcomes), instructional designers should be interested in the question regarding the effectiveness of the learning process. Therefore, effects of the presented three approaches should be investigated more closely. One should ask if subjective complexity really was lower (approach 1), if the material was designed to

help learning (e.g., the included picture enhanced the building of a correct mental model; approach 2), and whether activation of the learner's cognitive processes was successful (approach 3). To answer these questions, a differentiated investigation of invested cognitive resources during learning is needed. *Cognitive Load Theory* (CLT; Chandler and Sweller 1991, Sweller 2010a) presents a model describing different aspects of load that are relevant during learning. CLT assumes that working memory is limited in capacity, whereas long-term memory is limitless. The goal of learning in CLT is to construct domain-specific knowledge structures (i.e., cognitive schemas; van Merriënboer and Ayres 2005). But schema construction occurs in working memory with its limited capacity. Based on CLT, this capacity needs to be divided between different sources of load imposed on working memory during learning or dealing with different tasks. Traditionally, three independent sources of load, namely *intrinsic*, *extraneous* and *germane cognitive load* (e.g., Sweller et al. 1998) are differentiated. A refined view on CLT (Sweller et al. 2011) only separates intrinsic and extraneous cognitive load, and will be discussed later. The traditional view, from a practical perspective, helps in the investigation of the above-mentioned questions of instructional designers more clearly: (1) Investigating intrinsic cognitive load helps in understanding the subjective complexity placed on the learners, (2) investigating extraneous cognitive load helps in understanding the impact of differences in design, and (3) investigating germane cognitive load helps to understand the effort learners invested in learning. Therefore, in the following, the different concepts from the traditional view are described more in-depth (but can easily be adapted to the new view on CLT), used in the following studies, and used as a basis in the questionnaire to access cognitive load during learning.

Intrinsic cognitive load (ICL). The number of interacting elements (element interactivity) is the primary determinant of the amount of perceived intrinsic cognitive load (Sweller 1994; Sweller and Chandler 1994). For example, adding two random numbers no higher than 10 is less complex than adding three random numbers higher than 100. Therefore, adding the lower numbers results in less element interactivity, and theoretically, in less ICL imposed on a learner. Also, prior knowledge is an important factor for the perceived intrinsic load. Although experts are able to treat associated elements as one based on existing schemas, and are thus able to effectively augment their working memory capacity, novices cannot (Artino 2008). The combination of both prior knowledge (with respect to expertise of a learner) and nature (complexity) of the material therefore determine the perceived ICL of a learner (Sweller et al. 1998). To sum up, from an instructional design perspective, ICL can only be influenced by providing less or simpler learning content, or by providing some pre-training to activate prior knowledge and existing schemata. Therefore, investigating ICL more closely during learning is linked to the question of instructional designers, which content should be presented, and which prior knowledge of learners is assumed to effectively and successfully handle the given the task (approach 1).

Extraneous cognitive load (ECL). ECL results from the design of learning material. Therefore, to minimize ECL, anything that distracts the learner and hampers the learning process should be avoided in the design of learning material. High ECL results from all activities a learner needs to perform, even if they are not task relevant. This can be searching for relevant and masking/ignoring the irrelevant information, or the need to process an unnecessarily high number of elements simultaneously in working memory because of instructional design factors. High ECL therefore negatively affects the acquisition of knowledge. In instructional design research, this type of load has been especially addressed during the last few decades. A variety of multimedia design principles, effects, or strategies (for different overviews, see Sweller et al. 1998; Mayer 2009, 2005b; Plass et al. 2010) have been invented and investigated to identify the design needs of good learning material

and to reduce the unproductive load placed on a learner's working memory. Two of these principles are used in the presented studies, and are therefore described more closely: (1) The worked-examples effect (for an overview, see Atkinson et al. 2000; Renkl 2005) states that learning from an example is more effective than learning through problems: Providing learners with the steps of how to come to a correct solution, and the solution of a problem itself helps them to construct schemata of how to solve a problem, and avoids means-ends analyses. Also, activities connected to general search strategies—needed when provided with problems instead of worked examples—are unproductive for learning, and therefore produce unnecessary (which means extraneous) cognitive load (Renkl and Atkinson 2010). (2) The split-attention principle (Ayres and Sweller 2005, also known as contiguity principle: Moreno and Mayer 1999) states that it is important to physically and temporally integrate different sources of information to avoid unnecessary search processes during learning. When providing learners with multiple representations, mapping between various sources of information is necessary to construct an integrated mental model of these resources. Searching for corresponding elements in all parts of the learning material causes unproductive ECL and could be avoided by presenting information in an integrated format, so learners do not need to perform search processes because of instructional design factors. This means that the different sources of information should be spatially integrated (e.g., corresponding parts of the text are near the corresponding elements in the picture, or different pictures are arranged to represent their natural formation). Figure 1, which was used in Study 4 of this paper, is a good example of this. Concluding, from an instructional design perspective, investigating effects of design is very important for providing learners with effective learning material that enables learners to allocate their attention to task-relevant issues. Investigating ECL, therefore, is linked with the question of how instructional designers can design less demanding learning material (approach 2).

Germane cognitive load (GCL). As already mentioned, GCL is the load imposed on working memory that can have a positive impact on learning. Therefore, unlike the other two loads, GCL is a productive load that helps with schema acquisition and automation (Moreno and Park 2010). GCL arises within the learner during constructing, processing, and automating mental models or schemas. A possible strategy for learners to invest more into, for example, building a mental model, is to use learning strategies, such as, cognitive, metacognitive, or resource-oriented strategies. Directly connected to learning and building mental models are cognitive strategies, such as rehearsal strategies, organization of the learning material, and elaboration strategies, which can help one to gain deeper knowledge of the learning content. As a designer or creator of learning material, GCL can be fostered, for example, by including cognitive or metacognitive prompts, or by including desirable difficulties. Both instructional means have the potential to increase learners invested germane resources. (1) Prompts (Bannert 2009) can activate learning strategies and foster self-explanations (e.g. as done in Study 5 of this paper by encouraging the learner to imagine a picture of the content to be learned). This should enrich the mental model that originated from the written text. (2) Including desirable difficulties into the learning material, such as disfluency—which leads to a less fluent learning experience—to increase subjective difficulty (Alter and Oppenheimer 2009) may also foster learning. Whether this increase in learning can be seen as an effect of increased germane load is still under debate. However, we argue from the definition of germane load that learners invest increased effort into cognitive processes. In contrast to the direct effects of prompts, disfluency can be seen as a more subtle trigger, which nevertheless can have the same effects: Learners evaluate the learning material as more difficult and thus adjust their invested effort and the depth of their processing (Kühl and Eitel 2016). It should also be mentioned that the potential

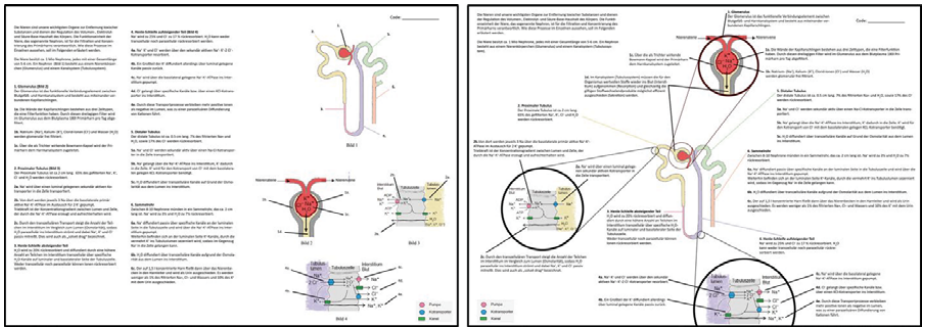

Fig. 1 Example of split-source (left) and integrated (right) format

positive trigger of disfluency might come along with the negative costs of increased extraneous load. Both effects, the increase in germane as well as in extraneous load have been demonstrated in Seufert et al. (2017), who analyzed increasing levels of illegible fonts in their experiment. They found that only in the worst legible condition ECL was affected. With standard variations of only one disfluent condition in contrast to a fluent condition as is also used by Eitel et al. (2014) no differences in ECL could be found. Therefore, with variations which are less legible than normal text but not too disturbing, ECL seems unaffected which – together with the assumed increase in GCL – could then lead to deeper processing during reading and thus to improved learning outcomes. Besides fostering GCL through instruction, motivation seems to be an important factor (Debue and van de Leemput 2014): GCL was found to be positively related with motivation. Higher intrinsic motivation resulted in a higher reported ability to devote cognitive resources to learning. To sum up for GCL, instructional design is interested in effective activation of the learner's cognitive processes (approach 3) to help with schema construction, automation, and building of enriched mental models to foster learning.

As already mentioned, the refined view on Cognitive Load Theory (Sweller et al. 2011) only separates intrinsic and extraneous cognitive load which derives from the features of the learning material. To handle these affordances, learners have to allocate their working memory resources in accordance: when they deal with the intrinsic affordances they invest germane resources, when they deal with extraneous affordances they invest extraneous resources. The allocated resources in working memory to deal with the extraneous cognitive load, do not need further assessment because they reflect nothing other than ECL in its former view (Kalyuga 2011). Allocation of germane resources for dealing with the imposed intrinsic cognitive load, needs to be examined more closely: The perceived intrinsic load reflects what learners perceive as the objective affordances (i.e. the complexity defined by the numbers of inter-related elements) in relation to their own capabilities to deal with the learning task (i.e. their prior knowledge). However, when it comes to germane resources, it could be that learners allocate much more germane resources than needed to deal with the imposed intrinsic load by the learning material itself. The additionally allocated working memory resources can then be used to foster understanding, for example, building mental models and schemata or automation by using for example different learning strategies like elaboration strategies, metacognitive strategies or the use of given prompts. This would reflect the original concept of germane cognitive load.

Overall, coming back to instructional design, instructors and designers of learning material need to know how learners process the learning material, how mentally loading these materials turn out to be, and whether activation of learners' cognitive processes is successful. Therefore, in the last few decades, measuring cognitive load during the learning process has become important. For a long time, cognitive load during the learning typically was measured by asking learners about their invested mental effort. This item was invented by Paas (1992), and is widely used in CLT research. Unfortunately, it is not possible to differentiate different aspects of load with one item. Therefore, a few researchers have made an attempt to find other measures of cognitive load. For an overview, see Brünken et al. (2010), or Brünken et al. (2003): There are some objective measures through physiological parameters (e.g., pupil dilation, brain activity), behavior (e.g., time-on-task), and performance (e.g., learning outcome, dual-task performance), as well as subjective ones, which, strictly speaking, are self-reports (e.g., invested mental effort, stress level). But all these approaches only allow one to assess the overall amount of load a learner experiences during learning.

Newer research has focused on the development to measure the different aspects of cognitive load separately. For an overview, see Klepsch et al. (2017). Whereas many approaches have not been investigated more closely, or have been developed to fit one instructional design research study, two approaches stand out: (1) Leppink et al. (2013) developed a questionnaire, which was enhanced in Leppink et al. (2014). Leppink et al. (2013) developed 10 items, three to measure ICL, three to measure ECL, and four to measure GCL. They found the three factors in their questionnaire to be robust, but mentioned that there is no evidence that these three factors represent the three types of cognitive load (Leppink et al. 2014). (2) Klepsch et al. (2017) also developed a questionnaire to measure cognitive load in a differentiated way. They also have been able to provide a questionnaire showing three factors: they used two items to measure ICL, three items to measure ECL, and three items to measure GCL. Based on their research design, they claim to actually measure differences in the types of loads, but mention that the questionnaire still must be validated in different and more ecologically valid learning contexts.

Besides both approaches needing more investigation, Leppink et al. (2014) stated that they are more in line with the proposed reconceptualization of CLT, whereas Klepsch et al. (2017) stuck to the traditional view of CLT. This, in our view, makes the questionnaire of Klepsch et al. (2017) more helpful when one, like us in this paper, is interested in instructional design effects and their impact on the learning task itself (approach 1), the design of the material (approach 2), and the activation of the learner's cognitive processes (approach 3). To help in theory building of CLT, both questionnaires should be considered in further research.

In this paper, we were especially interested in the effects of instructional design and their explanation through CLT, and therefore the three-factor structure of CLT because it is more practically useful when dealing with instructional design. Therefore, we needed a measurement method of CLT that, in a first step, differentiates between different types of load, and in a second step, could be used in various learning situations and with a wide variety of learning materials. As adoption of the questions with the questionnaire of Leppink et al. (2013) is more complex, and therefore might lack comparability over different topics we used the questionnaire of Klepsch et al. (2017). In their paper they found evidence for the factor structure with the three types of load as well as satisfying reliability scores for the questionnaire. However, they validated the questionnaire by using rather short and artificial learning tasks. Thus, with the current paper we want to add the missing element on validating the questionnaire by using real life learning settings. Therefore, we review a set

of different studies in natural learning settings. The six studies use different approaches to instructional design that are linked to concepts in CLT: (1) Differences in element interactivity (Study 1) or variations of difficulty based on knowledge (Study 2) are used to manipulate intrinsic cognitive load by varying the learning task itself. (2) By manipulating the design through the worked example principle (Study 3) and the split attention principle (Study 4) different levels of extraneous cognitive load were triggered allowing us to investigate the influence of different designs on ECL. (3) Variations in prompted imagination of written learning material (Study 5), as well as differences in legibility of learning material (Study 6), are used to directly or indirectly activate the learner's cognitive processes, and therefore induce germane cognitive load, to create or foster enriched mental models.

Our goal is to substantiate the sensitivity of the questionnaire by Klepsch et al. (2017) for instructional design-based variations in cognitive load. We want to find out if the questionnaire is able to differentiate exactly the type of load that has been intentionally varied based on theoretical assumptions and previous empirical studies. This would be a strong argument for its prognostic construct validity. In addition, we will also add some information on the reliability of the questionnaire by Klepsch et al. (2017) based on our six studies.

Hypotheses

H1a Concerning ICL, we expect no difference in Studies 3, 4, 5, and 6 between the experimental conditions of each study because learning content has not been varied in element interactivity, complexity, or quantity.

H1b Concerning ICL, based on CLT and the assumptions about element interactivity and prior knowledge, we expect that ICL varies significantly between the three conditions in Studies 1 and 2 because in these studies, ICL was varied on purpose: We expect the participants in the boredom conditions to report the lowest ICL, and the participants in the overload condition to report the highest ICL.

H2a Concerning ECL, we expect no significant differences in Studies 1, 2, 5, and 6, because the design of the learning material has not been varied.

H2b Concerning ECL, the design of the learning material was varied on purpose in Studies 3 and 4, therefore we expect significant differences between the two conditions of each study.

H3a Concerning GCL, we do not expect differences between the experimental conditions in Studies 1, 2, 3, and 4, because we did not include any direct or indirect activation of the learner's cognitive processes.

H3b Concerning GCL, we expect significant differences between the experimental groups for Studies 5 and 6 because either activation of the learner's cognitive processes was directly implemented (prompts) or indirectly triggered (legibility of learning material).

H4 Concerning learning outcomes, we expect a significant difference in task performance in the experimental groups for all studies. In Study 1 and 2, task performance in

the boredom group should be highest where ICL is lowest, whereas task performance in the overload condition will be lowest where ICL is highest. In Study 3 and 4, task performance should be highest in the groups where multimedia principles are applied correctly to help learning, where ECL should be lowest. In Study 5 and 6, task performance should be highest when schema construction was activated, or deeper processing was induced, and therefore GCL should be highest.

Method

In the following, 6 studies are described: two of them varying ICL, two of them ECL, and two of them GCL. All the studies follow the same methodological approach of varying one type of load on purpose, and measuring the effects of this variation on all three types of load, as well as on learning outcomes. Thus, there are many similarities which will be presented in the following for all studies together. In the subsequent paragraph, the differing key features of each study (e.g., the operationalization in the load variation) will be presented. For an overview of all six studies, see Table 1.

Participants

Each study had between 31 and 62 participants, with at least 15 participants in each experimental condition. Table 1 provides an overview of the distribution between the sexes and average age for each study, and its experimental conditions.

Design

All studies followed the same experimental design principle: They are all between-subject studies with either two (Study 3, 4, 5, and 6) or three (Study 1 and 2) experimental conditions. Independent variables are the variations of one type of load each, which has been operationalized in different experimental conditions for each study. In Study 1 and 2, where ICL has been varied, the conditions vary in task difficulty. In Study 3 and 4, where ECL has been varied, the conditions varied in the design features of the learning material. In Study 5 and 6, where GCL has been varied, the conditions differ in features that should activate the learner's cognitive processes. All these variations should result in measurable differences of the dependent variables (i.e., in either intrinsic, extraneous, or germane cognitive load), as well as in learning outcome (Study 3 to 6) or task performance (Study 1 and 2).

As control variables, age and sex have been assessed in each study. For Studies 3 to 6, prior knowledge also has been assessed. There was no prior knowledge assessed in the studies varying ICL because the tasks we used were adopted from flow research and were programmed to adopt to participants prior knowledge and their skills. While playing Tetris (Study 1) the program stayed either way below participants' skill, adopted to their skill or was way too difficult and exceeded their skill. Thus, prior knowledge in this case would reflect prior experience, coming along with the skill to fluently react on the given task. Nevertheless, we did not measure it directly because it was implemented in the

Table 1 List of studies and their assignment to a type of load, number of participants in each study and each experimental group, as well as mean age and sex distribution across the six studies

	Nr	Learning domain/task	N	M_{age} (SD_{age})	%male	Experimental groups	n
ICL	1	Problem solving: Tetris game	62	25.30 (9.56)	65.6%	Easy gameplay (boredom)	21
						Adequate gameplay (flow)	20
						Difficult gameplay (overload)	21
	2	General knowledge: Quiz	62	25.21 (9.51)	66.1%	Easy questions (boredom)	20
						Adequate questions (flow)	22
						Difficult questions (overload)	20
ECL	3	Mathematics: extremum problems and Taylor polynomial tasks	40	22.78 (5.14)	17.5%	Problem solving	20
						Worked examples	20
	4	Biology: kidney	51	23.16 (4.28)	7.8%	Split source format	26
						Integrated format	25
GCL	5	Technology: clutch	31	28.39 (9.99)	45.2%	No imagination	15
						Imagination	16
	6	Geography: Earth's time zones	36	21.44 (3.74)	8.3%	Fluent	18
						Disfluent	18

program's logical structure of referring to the learners' skills. For answering questions in "Who Wants to be a Millionaire" (Study 2) we referred to learners' general knowledge as prior knowledge, which we didn't measure directly either. Instead, we ensured that learners had sufficient educational background with the "Abitur", a diploma from German secondary school qualifying for university admission or matriculation. In Studies 5 and 6, in which the induction of GCL is varied, we also assessed learners' current motivational state (FAM; Rheinberg et al. 2001).

Procedure

The procedure was also the same for all the studies. In the beginning, learners were informed about the procedure of the study they participated in and signed an informed consent form. All participants were aware that they could withdraw their data at any point in the study without having any disadvantage. Then, each participant was randomly assigned to one of the experimental groups. As a next step in each study, each participant filled out a questionnaire asking for demographic data.

Then learners' prior knowledge and motivation were assessed, whenever needed, as described above. Afterwards, participants had to deal with the task or the learning material as described separately for each study in section "3.4 Material".

After learning or performing the task, the cognitive load questionnaire (Klepsch et al. 2017) had to be answered, followed by a knowledge test for the corresponding study in Studies 3 to 6. Performance in Study 1 and 2 was calculated directly through task performance by the used programs (see "Material"), and in these cases, only the cognitive load questionnaire had to be answered after performing the given task.

Material

In the following six paragraphs, the used learning materials, the characteristics of the experimental variations, and the specific features of the learning task are described for each study. Also, additionally included questionnaires are described. In a 7th paragraph, the questionnaire for measuring the three types of load is described in more depth.

Study 1 (varying ICL): problem solving – Tetris game.

The Tetris game (Keller and Bless 2008) used was originally designed for research on flow experiences (Keller and Bless 2008). We used it to vary intrinsic cognitive load. The three conditions, (1) boredom, (2) flow, and (3) overload, varied in element interactivity in relation to the learner's skills by changing complexity of the task as an independent variable. In the boredom condition, blocks came down slowly and the participants had plenty of time to position them in the right place. In the flow condition, the program adapted itself to the performance of the participant. If one produced fewer mistakes, the level increased and therefore the game became more difficult, if one made mistakes, the game got easier. The overload condition was programmed to start on a high level and did not reduce the level based on performance. Participants were instructed to create as many full lines as possible and they played the game for three minutes. Despite learning to play Tetris or practicing Tetris is not related to knowledge acquisition it nevertheless is a task where information has to be processed in working memory to reach a certain goal and to refine their strategies in solving this task. As the information (blocks) interact with each other it can be seen as an operationalization of element-interactivity which is a key feature of intrinsic load. Time was restricted, as otherwise participants would have the possibility to try many different strategies, and we did not like to induce germane load. Based on the given instruction participants dealt with a problem-solving task. Through the log files of the program, the number of blocks a participant got during the game, and the number of finished lines was documented during the three minutes of game play. Therefore, task performance could be estimated as followed: $((\text{finished lines} \times 10) / (\text{received blocks} \times 4)) \times 100$.

Study 2 (varying ICL): general knowledge – quiz.

The Knowledge Quiz (Keller et al. 2011) was a computerized quiz based on questions from the German version of the board game "Who Wants to be a Millionaire?" (Jumbo Spiele® 2000), which is based on a TV show equally named. Each question has four possible answers, and only one is correct. Participants played the game for three minutes, and they had to answer each question within a limited time period. Similar to Study 1 participants were randomly assigned to one of three conditions: (1) boredom, (2) flow, and (3) overload. This conditions varied ICL as independent variable by varying the difficulty of the presented questions. In the boredom condition, questions were easy and stayed at this level. In the flow condition, difficulty of the level was adapted to the skills of each participant. If one could handle a certain number of tasks in one level, the level went up and questions became more difficult. If one failed to answer a certain number of questions right, the level went down. As a result, the game always adjusted to fit to the skills of the participant. In the overload condition, the questions were so difficult that participants most times only had

the option to guess. Task performance was converted into percentage of correct answers based on the individually reached number of questions, which varied based on the pace in which questions were answered.

Study 3 (varying ECL): mathematics – extremum problems and Taylor polynomial tasks.

Each learner had to deal with two types of mathematical learning material: (1) extremum problems and (2) Taylor polynomial tasks. Prior knowledge was assessed after participants received an example task for each type of material: in two questions, they had to answer on a 7-point Likert-type scale whether they were familiar with such tasks, and in a second step, they had to write down how they would solve the task. Afterwards, both learning materials started with a short introduction on the topic, followed by two example tasks. To realize differences in ECL as an independent variable, (1) problem solving or (2) worked examples were included. In the problem-solving group, the right solution was presented, but how the solution could be calculated was missing. In the worked example group, the process of calculating the right answer for the example tasks was written down. Presentation of the two materials was randomized. Participants had 5 min time for each domain to learn the content, and another 15 min to do the post test. Performance was measured by a post test for each domain containing two analogue tasks (similar to the example tasks in the learning material), and one complex transfer task. For each task, 6 points could be reached, and therefore as a maximum 36 points could be reached in the post test. Task Performance was calculated by summing up points given for correct answers and converted into percentage.

Study 4 (varying ECL): biology – kidney.

Prior knowledge about the human kidney was assessed through 10 open questions, covering content of the following learning material. The material about the structure and functions of the kidney itself consisted of four pictures and seven corresponding texts. As a dependent variable, ECL was varied through either (1) split source material or (2) integrated material. In the split attention group, the text was on the left side of the paper and the four pictures on the right side. References, in the form of numbers, connected the pictures with the text. In the integrated format, the interplay between the four pictures was made clear and each text was connected within the corresponding part in the pictures. Through this approach, search processes between text and pictures could be minimized. Each participant had 30 min to learn the content. To assess learning performance, a post test had to be filled in consisting of 13 tasks. Altogether, 35 points could be reached. Individual performance was converted into a percentage.

Study 5 (varying GCL): technology – clutch.

Prior knowledge was assessed through eight questions, including multiple choice questions, open questions, and a task asking to draw a clutch. The learning material about clutches in cars itself consisted of three parts, with overall 682 words. Instruction on imagination of the learning content varied between groups because GCL was the independent variable: Both groups received the instruction to learn the text and were allowed to make notes during learning, but only one group was instructed to imagine a picture of the described clutch

and its functions. The time needed by each learner to work through the learning material was measured. To assess learning performance, a post test was conducted consisting of 14 questions, with at most 39 points. Performance was recalculated into percent. Before the learning phase in this study, current motivational state with the subscales anxiety of failure, probability of success, interest, and challenge (FAM; Rheinberg et al. 2001) was additionally assessed.

Study 6 (varying GCL): geography – Earth's time zones.

Seufert et al. (2017) already reported a study on text legibility and allowed us to use their data. Prior knowledge about the Earth's times zones was assessed through six tasks on time differences, which were either open or multiple-choice tasks. The learning material was a text about the Earth's time zones (adapted from Schnotz and Bannert 1999). It included 1,100 words and a table with time differences for eight cities. To vary GCL as an independent variable, the texts legibility was either (1) fluent (Arial, black 12 pt) or (2) disfluent (Monotype Corsiva, black, 12 pt). Time for learning was not restricted, and therefore measured (about 13 min: M = 763 s, SD = 227.51). Learning performance was measured through a post-test with 17 questions, where, at most, 34 points could be reached. Again, performance was converted to percent to foster comprehension of the six studies. For this study, learners' current motivational state (FAM) was also measured.

Cognitive load questionnaire (CLQ).

To measure cognitive load differentially, the cognitive load questionnaire (CLQ) by Klepsch et al. (2017) was used. This questionnaire has eight items to measure the different aspects of cognitive load. The eight items disperse on three scales. A scale with two items asking for perceived ICL, a scale with three items asking for perceived ECL, and finally a scale of three items asking for perceived GCL. The GCL scale includes an item on "elements that helped one to understand the content," which directly addresses, for example, something like prompts. Klepsch et al (2017) argued that this item could be excluded if there are no such elements in the learning material, and thus would decrease internal consistency. Based on theoretical assumptions, we would assume that in studies 1–4 and 6, reliability should be higher when leaving out the third GCL item. For Study 5, we assume that reliability should be best if all items are included. To demonstrate this, in each study, all items of the questionnaire had to be answered. All items must be rated on 7-point Likert-type scales ranging from "*completely wrong*" to "*absolutely right*". For our studies, the items are presented in German, the scale's original language (see Appendix 1 for the English and German version).

Data analysis

Data were analyzed in the same way for all six studies: For all three scales of the differentiated questionnaire on Cognitive Load, reliability was estimated by using McDonald's ω. To provide evidence for the already described issue, with the third item of the GCL scale, the scale on GCL was analyzed with the third item (GCL3) and without (GCL2). A confirmatory factor analysis was conducted following Klepsch et al. (2017). We analyzed the data for two models: Model 1 assumes that one unitary factor accounts for all load items. Model

2 represents the theoretical assumptions of three inter-related types of cognitive load and suggests the three-factor model with separable but related factors. Adequate model fit is indicated by a low chi-square (χ^2) value, a high Tucker-Lewis index (TLI $\geq$ 0.95), a high comparative fit index (CFI $\geq$ 0.95), and a low root-mean-square error of approximation (RMSEA $\leq$ 0.05). Additionally, the Akaike information criterion (AIC) was used as a model comparison index: The model yielding the lowest AIC is preferred in terms of close model fit and parsimony of the model relative to competing models.

The variables on prior knowledge, age, and, if accessed, other control variables (as described in part 3.1 Design, Materials and Procedure) have been tested for group differences to check for differences in randomization. Also, correlations between these variables and the dependent variables have been calculated. Whenever a significant difference between groups or a significant correlation could be found, the affected variable was included in the following analyses of variance as a covariate. All correlations tables can be found in Appendix 2.

To identify group differences for all studies one-way AN(C)OVAS were conducted to enable comparability of the six studies. In studies with three conditions (Study 1 and 2), contrasts were also reported to show differences between the experimental groups. Whenever multiple contrast analyses have been conducted, we adjusted the alpha level.

Results

McDonalds ω was conducted to assess the reliability of the three scales of the differentiated cognitive load questionnaire. For ICL, two items, and for ECL, three items were included. The analysis for GCL was conducted with two items. Only for Study 5, as recommended by Klepsch et al. (2017), we also conducted an analysis for GCL with three items as the prompts in the material correspond to the third GCL-question. McDonalds ω for all scales in all studies is reported in Table 2.

The confirmatory factor analysis was conducted over all studies. All participants of all studies were included in one overall analysis. Model fits can be found in Table 3. Based on the sample size a multilevel or multigroup analysis, as it has been conducted in Klepsch et al. (2017) was not possible, as there are not enough participants in the different studies or conditions. Model 1 assumes one unitary factor and its model fit is rather bad. Model 2 represents the theoretical assumptions of three inter-related types of cognitive load (three factor model with separable but related factors). Model fit is not perfect as the chi-square (χ^2) value is still quite large, the Tucker–Lewis index is lower than 0.95 the root-mean-square error of approximation is higher than 0.05.

In the following six sections, differences in the experimental groups of each study are reported using a one-way ANOVA or ANCOVA, if covariates (age or prior knowledge) are included. For Study 1 and 2 contrasts are also reported because these studies include three experimental groups.

Study 1 (varying ICL): problem solving – Tetris game

To repeat: In this study three experimental groups were compared by varying ICL while playing Tetris: (1) The boredom condition, where the Tetris game was adjusted to be simple (−> low ICL), (2) the flow condition, where the game adapted to learners' performance

Table 2 McDonalds's ω for each scale in each study

	Nr	Learning content/task	McDonalds ω		
			ICL	ECL	GCL
ICL	1	Play Tetris	.87	.75	.72
	2	Answer knowledge questions	.85	.72	.75
ECL	3	Solve mathematical problems	.83	.81	.74
	4	Learn about the kidney	.80	.85	.72
GCL	5	Learn about the clutch	.87	.77	2 Items: .84 3 Items: 64
	6	Earth's time zones	.87	.79	.70

Table 3 Fit indices for Model 1 and Model 2

	$\chi 2$	*df*	*p*	TLI	CFI	RMSEA	AIC
Model 1	264.10	14	< .001	.38	.69	.25	306.10
Model 2	51.84	13	< .001	.90	.95	.10	95.84

(− > medium ICL), and (3) the overload condition, where the game was quite difficult (− > high ICL).

For ICL, as expected, we found a significant main effect ($F(2,59) = 10.01$, $p < 0.001$, $\eta^2 = 0.25$): Contrasts showed that ICL was lowest for participants in the boredom condition. They reported significantly lower ICL than participants in the flow ($T(59) = 3.51$, $p < 0.01$, $d = 1.25$) or overload ($T(59) = 4.15$, $p < 0.001$, $d = 1{,}19$) condition. We found no difference between the flow and overload condition ($T(59) = 0.59$, $p = 0.56$, $d = 0.18$).

For ECL, we found a significant main effect ($F(2,59) = 12.19$, $p < 0.001$, $\eta^2 = 0.29$) that was not expected. Contrasts showed that ECL was lowest for participants in the boredom condition. They reported significantly lower ECL than did participants in the flow ($T(59) = 3.85$, $p < 0.01$, $d = 1.13$) or overload ($T(59) = 4{,}59$, $p < 0.001$, $d = 1.74$) conditions. We found no difference between the flow and overload condition ($T(59) = 0.69$, $p = 0.49$, $d = 0.19$).

For GCL, as expected, we found no significant main effect ($F(2,59) = 2.86$, $p = 0.07$, $\eta^2 = 0.09$), nor could we find any significant contrasts (boredom-flow: $T(59) = 1.93$, $p = 0.06$, $d = 0.75$; boredom-overload: $T(59) = 1{,}93$, $p = 0.06$, $d = 0.50$; flow-overload: $T(59) = 0.18$, $p = 0.99$, $d = 0.23$).

For task performance, we found a significant main effect ($F(2,59) = 86.04$, $p < 0.001$, $\eta^2 = 0.75$): Contrasts showed that the participants in the boredom condition scored significantly higher than participants in the flow condition ($T(57) = 6.16$, $p < 0.001$, $d = 0.43$) and participants in the overload condition ($T(57) = 13.11$, $p < 0.001$, $d = 3.94$). Also, participants in the flow condition scored higher than participants in the overload condition ($T(57) = 6.95$, $p < 0.001$, $d = 3.05$).

All means and standard deviations can be found in Table 4, and effects can be found in Fig. 2.

Table 4 Means and standard deviation of all dependent variables in Study 1

Study 1	Boredom		Flow		Overload	
	M	SD	M	SD	M	SD
ICL	2.14	1.42	3.7	1.20	4.05	1.77
ECL	2.13	0.93	3.63	1.64	3.90	1.10
GCL	2.76	1.59	4.00	1.71	3.60	1.78
Task performance	69.26	19.24	42.44	12.23	12.21	6.96

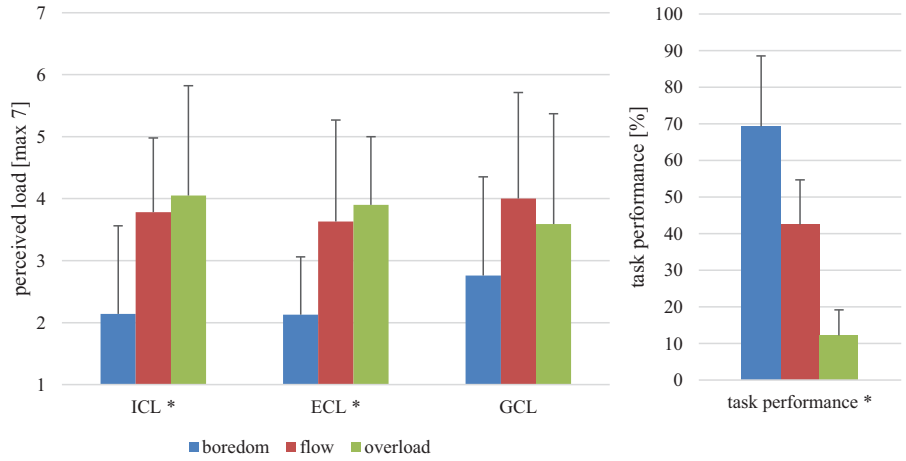

Fig. 2 Effects on the dependent variables for Study 1. *Note* *main effect with $p < .05$

Study 2 (varying ICL): knowledge quiz

In this study, ICL was also varied. Three experimental groups were answering quiz questions with varying difficulty: (1) in the boredom condition, quiz questions where simple (−>low ICL), (2) in the flow condition, questions adopted to learners' knowledge (−>medium ICL), and (3) in the overload condition, questions were too difficult to know the answer (−>high ICL).

For ICL, as expected, we found a significant main effect ($F(2,59) = 8.76$, $p < 0.001$, $\eta^2 = 0.23$): Contrasts showed that ICL was lowest for participants in the boredom condition. They reported significantly lower ICL than participants in the flow ($T(59) = 3.40$, $p < 0.01$, $d = 1.24$) or overload ($T(59) = 3.84$, $p < 0.001$, $d = 1.32$) conditions. We found no difference between the flow and overload conditions ($T(59) = 0.54$, $p = 0.59$, $d = 0.14$).

For ECL, we found no main effect ($F(2,59) = 2.79$, $p = 0.07$, $\eta^2 = 0.09$): Contrasts showed no differences between participants in the boredom and flow conditions ($T(59) = 0.94$, $p = 0.35$, $d = 0.34$) or in the boredom and overload conditions ($T(59) = 1,38$, $p = 0.17$, $d = 0.43$). But there was a significant difference between the flow and overload conditions ($T(59) = 2.35$, $p < 0.02$, $d = 0.66$).

For GCL, as expected, we found no significant main effect ($F(2,59) = 0.21$, $p = 0.81$, $\eta^2 = 0.01$), nor could we find any significant contrasts (boredom-flow: $T(59) = 0.07$, $p = 0.94$, $d = 0.02$; boredom-overload: $T(59) = 0,59$, $p = 0.56$, $d = 0.21$; flow-overload: $T(59) = 0.53$, $p = 0.60$, $d = 0.14$).

Table 5 Means and standard deviation of all dependent variables in Study 2

Study 2	Boredom		Flow		Overload	
	M	SD	M	SD	M	SD
ICL	1.25	0.41	2.55	1.39	2.75	1.56
ECL	3.18	0.93	2.80	1.28	3.75	1.61
GCL	3.10	1.07	3.14	1.94	3.40	1.67
Task performance	97.45	3.44	58.42	7.53	30.24	5.16

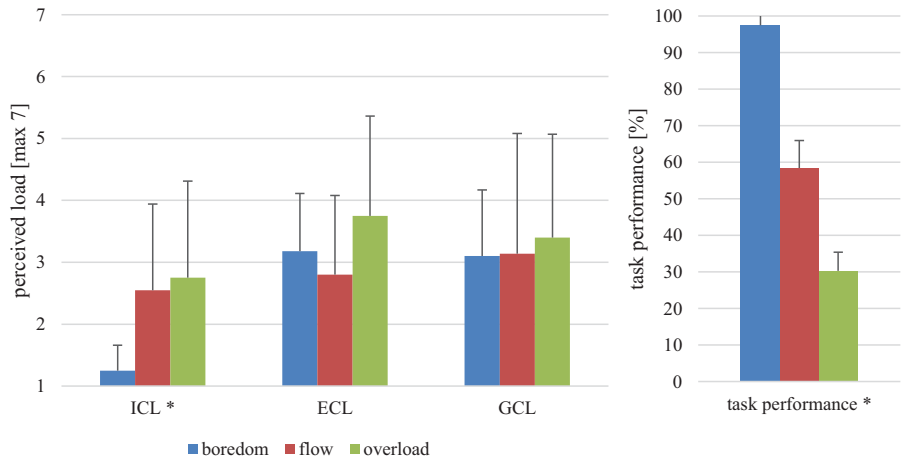

Fig. 3 Effects on the dependent variables for Study 2. *Note* *main effect with $p < .05$

For task performance, we found a significant main effect ($F(2,54) = 585.56$, $p < 0.001$, $\eta^2 = 0.96$): Contrasts showed that the participants in the boredom condition scored significantly higher than participants in the flow condition ($T(54) = 20.74$, $p < 0.001$, $d = 6.56$) and participants in the overload condition ($T(54) = 34.11$, $p < 0.001$, $d = 15.33$). Also, participants in the flow condition scored higher than participants in the overload condition ($T(54) = 15.22$, $p < 0.001$, $d = 4.33$).

All means and standard deviations can be found in Table 5, and effects can be found in Fig. 3.

Study 3 (varying ECL): mathematic – extremum problems and Taylor polynomial tasks

Keep in mind, in this study, two experimental groups are compared varying ECL: (1) The problem-solving condition, where only solutions are presented for example tasks (− > high ECL), and (2) the worked examples condition, where approach and solution have been presented for sample tasks (− > low ECL).

For ICL, as expected, we found no significant main effect ($F(1,38) = 0.09$, $p = 0.77$, $\eta^2 = 0.23$): ICL was reported to be equal in the experimental groups.

For ECL, as expected, we found a main effect ($F(1,38)=8.91$, $p<0.01$ $\eta^2=0.19$): In the worked examples group, ECL was reported to be lower than in the group without worked examples.

For GCL, as expected, we found no significant main effect for GCL ($F(1,38)=1.18$, $p=0.28$, $\eta^2=0.03$): Both groups reported similar GCL ratings.

For task performance, age was included as covariate into the analysis because we found a significant correlation between these two variables ($r=-0.38$, $p<0.05$). For task performance, we found a significant main effect ($F(1,37)=9.44$, $p<0.01$, $\eta^2=0.20$): In the worked examples group, task performance was significantly higher than in the group without worked examples.

All means and standard deviations can be found in Table 6, and effects can be found in Fig. 4.

Study 4 (varying ECL): biology – kidney

During this study, two experimental groups are compared varying ECL: (1) The split-source condition had separated representations of text and pictures (−>high ECL), and (2) the integrated format condition worked with an integrated version of text and pictures (−>low ECL).

For ICL, as expected, we found no significant main effect ($F(1,49)=3.33$, $p=0.07$, $\eta^2=0.06$): ICL was reported to be equal in the two experimental groups.

For ECL, age was included as covariate into the analysis because we found a significant correlation between these two variables ($r=-0.42$, $p<0.01$). For ECL, as expected, we found a main effect ($F(1,48)=9.60$, $p<0.01$ $\eta^2=0.17$): In the integrated format group, ECL was reported to be significantly lower than in the split source group.

For GCL, prior knowledge was included as covariate in the analysis because we found a significant correlation between these two variables ($r=0.29$, $p<0.05$). For GCL, as expected, we found no significant main effect for GCL ($F<1$ n.s.): Both groups reported similar GCL ratings.

For task performance, prior knowledge was included as covariate into the analysis because we found a significant correlation between these two variables ($r=0.43$, $p<0.01$). For task performance, we found a significant main effect ($F(1,48)=5.83$, $p<0.05$, $\eta^2=0.11$): In the integrated group, task performance was significantly higher than in the split-source group.

All means and standard deviations can be found in Table 7, and effects can be found in Fig. 5.

Table 6 Means and standard deviation of all dependent variables in Study 3

Study 3	No worked examples		Worked examples	
	M	SD	M	SD
ICL	4.39	0.93	4.28	1.41
ECL	3.67	0.98	2.71	1.04
GCL	5.54	0.74	5.28	0.79
Task performance	45.97	27.23	67.50	23.81

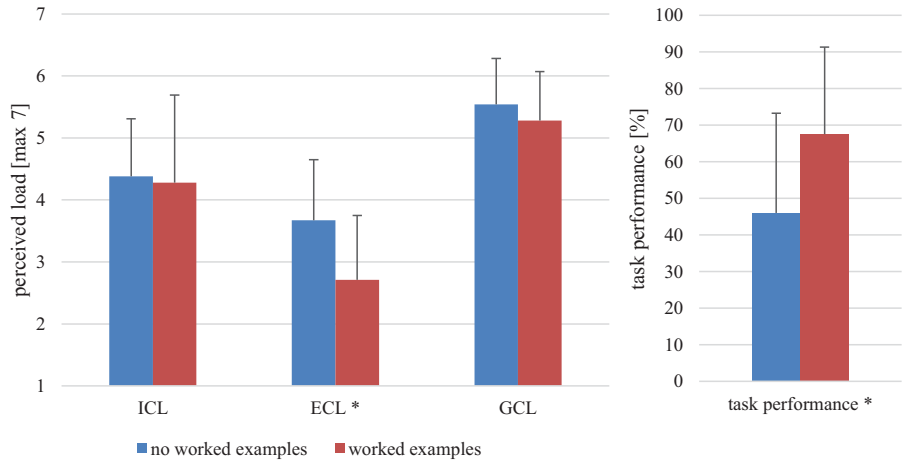

Fig. 4 Effects on the dependent variables for Study 3. *Note* * main effect with $p < .05$

Study 5 (varying GCL): technology – clutch.

Remember that in this study, two experimental groups are compared varying GCL while learning about clutches: (1) The group without a special instruction on how to handle the learning content (− > low GCL), and (2) the group with the instruction to imagine the learning content as a picture (− > high GCL).

In this study, the subscale on anxiety of failure of the questionnaire on actual motivation (FAM) was included in each analysis as a covariate because we found a significant difference between groups ($T(27) = 3.21$, $p < 0.01$, $d = 1.31$).

For ICL, additionally, prior knowledge and time on task were included in the analysis as covariates because we found significant correlations between these variables and ICL (prior knowledge: $r = -0.65$, $p < 0.01$; time on task: $r = 0.42$, $p < 0.05$). For ICL, as expected, we found no significant main effect ($F < 1$, n.s.): ICL was reported to be equal in both experimental groups.

For ECL, additionally, the subscales interest and challenge of the questionnaire on actual motivation (FAM) were included in the analysis as covariates because we found significant correlations between these variables and ECL (interest: $r = -0.40$, $p < 0.05$; challenge: $r = 0.45$, $p < 0.05$). For ECL, as expected, we found no significant main effect ($F < 1$, n.s.): Both experimental groups rated ECL similarly.

For GCL, additionally, time on task and the subscale interest of the questionnaire on actual motivation (FAM) were included in the analysis as covariates because we

Table 7 Means and standard deviation of all dependent variables in Study 4

Study 4	Split source		Integrated	
	M	SD	M	SD
ICL	6.04	0.86	5.50	1.22
ECL	4.53	1.58	3.44	1.19
GCL	5.50	1.21	5.90	0.95
Task performance	44.23	16.81	56.91	14.66

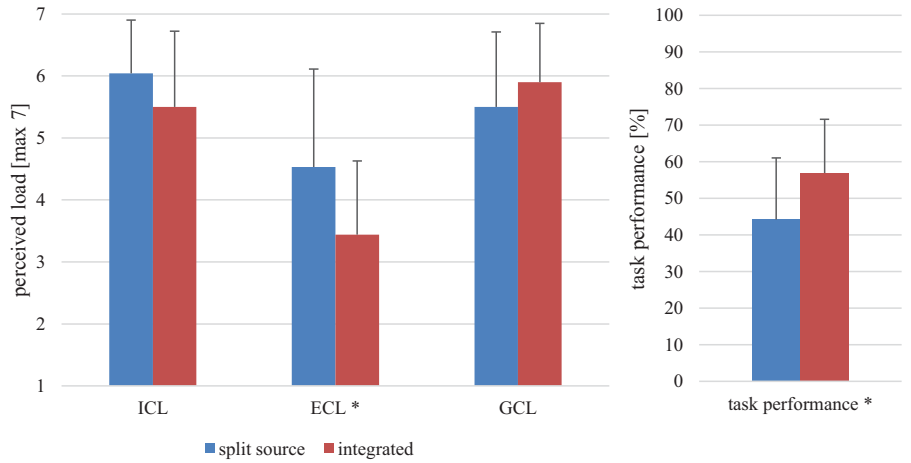

Fig. 5 Effects on the dependent variables for Study 4. *Note* * main effect with $p < .05$

found a significant correlation between these variables and GCL (time on task: $r = 0.26$, $p < 0.05$; interest: $r = 0.27$, $p > 0.05$). For GCL, as expected, we found a significant main effect ($F(1,51) = 3.14$, $p_1 = 0.03$, $\eta^2 = 0.07$): The group that was instructed to imagine the learning content reported higher GCL than the group that was not instructed to imagine the learning content.

For task performance, additionally, prior knowledge and the subscales interest and challenge of the questionnaire on actual motivation (FAM) were included in the analysis as covariates because we found a significant correlation between these variables and task performance (prior knowledge: $r = 0.40$, $p < 0.01$; interest: $r = 0.40$, $p > 0.01$; challenge: $r = 0.42$, $p < 0.01$). For task performance, we found no main effect ($F < 1$, n.s.): Both groups reached similar performance levels.

All means and standard deviations can be found in Table 8, and effects can be found in Fig. 6.

Study 6 (varying GCL): geography – earth's time zones

In this study, we compared two experimental groups varying GCL: (1) The fluent condition, with highly legible material (− > low GCL), and (2) the disfluent condition, with less legible material, which increases the subjective difficulty of the material (− > high GCL).

In this study, prior knowledge was included in each analysis as a covariate because, we found a significant correlation with each dependent variable (ICL: $r = -0.51$, $p < 0.01$; ECL: $r = -34$, $p < .05$; GCL: $r = 35$, $p < 0.05$; task performance: $r = 0.49$, $p < 0.01$).

For ICL, as expected, we found no significant main effect ($F(1,33) = 1.80$, $p = 0.18$, $\eta^2 = 0.05$): ICL was reported to be equal in the two experimental groups.

For ECL, probability of success was additionally included in the analysis as a covariate because we found a significant correlation between the two variables ($r = -0.43$,

Table 8 Means and standard deviation of all dependent variables in Study 5

Study 5	No imagination		Imagination	
	M	SD	M	SD
ICL	4.65	1.25	4.34	1.81
ECL	3.72	1.08	3.38	1.58
GCL	5.27	1.09	5.81	1.03
Task performance	54.04	25.78	59.13	20.65

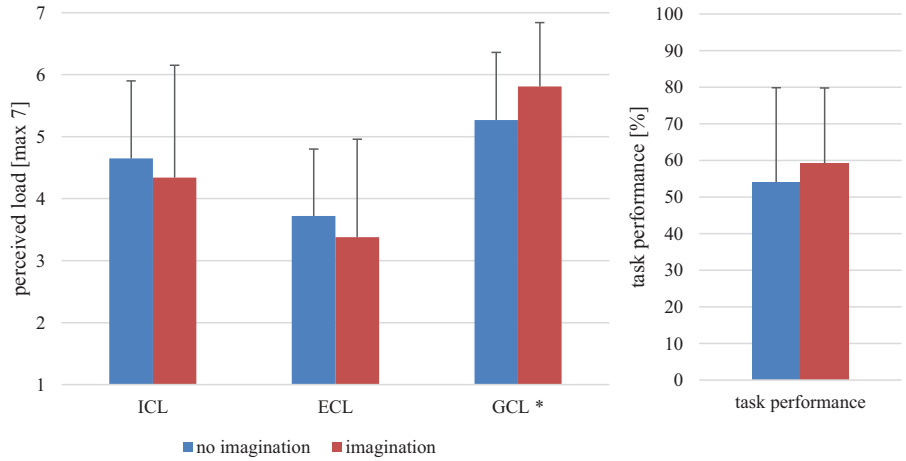

Fig. 6 Effects on the dependent variables for Study 5. *Note* *main effect with $p < .05$

$p < .01$). For ECL, as expected, we found no significant main effect ($F < 1$, n.s.): Both experimental groups rated ECL similarly.

For GCL, as expected, we found a significant main effect ($F(1,33) = 3.36$, $p_1 = 0.04$, $\eta^2 = 0.09$): Learners who received the disfluent learning material reported higher levels of GCL compared to learners who received the fluent learning material.

For task performance, we found no main effect ($F(1,33) = 1.65$, $p = 0.21$, $\eta^2 = 0.05$): Both groups reached similar performance levels.

All means and standard deviations can be found in Table 9, and effects can be found in Fig. 7.

Discussion

In a first step, we discuss the results of our six studies based on hypotheses (see "Validity"). Then we address the question regarding which of the items of the GCL scale are necessary in terms of internal consistency and whether a confirmatory factory analysis reveals similar results as found by Klepsch et al. (2017) (see "Reliability and Confirmatory Factor Analysis"). Also, strengths and weaknesses of the six-study approach in general are discussed (see "Strengths and Weaknesses"), and further directions are addressed (see "Further directions"). In a second step, we discuss the usefulness of a differentiated

Table 9 Means and standard deviation of all dependent variables in Study 6

Study 6	Fluent		Disfluent	
	M	SD	M	SD
ICL	4.53	1.34	4.28	1.44
ECL	3.06	1.20	2.94	1.03
GCL	5.33	1.32	5.78	0.73
Task performance	62.50	22.14	66.42	23.71

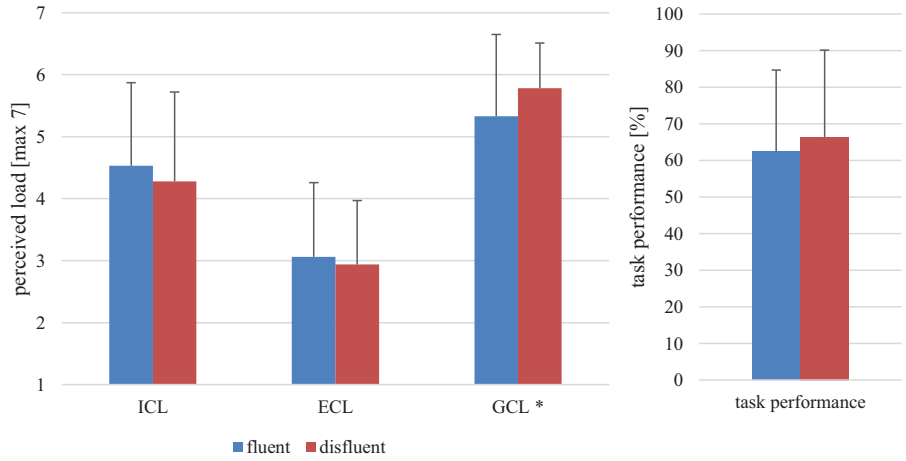

Fig. 7 Effects on the dependent variables for Study 6. *Note* *main effect with $p < .05$

measurement of cognitive load, especially the theoretical and practical implications and the need for further research (see "Conclusion").

Reliability and confirmatory factor analysis

The first criterion of a good measurement is its reliability. For ICL we found the scale to be robust in reliability with ω's higher than 0.80 (all studies), which is in line with the results found by Klepsch et al. (2017). Reliability of the ECL and GCL scale was lower but still acceptable for experimental field studies with ω's higher than 0.70. The GCL scale was calculated with two items (leaving out the item "The learning task consisted of elements supporting my comprehension of the task") only for Study 5 we in addition calculated it with 3 items This was based on theoretical assumptions: We expected that the item should best be excluded in Studies 1–4 and 6 because they do not comprise any elements that explicitly foster learners' germane load investment. Only Study 5 contained an element that theoretically is assumed to be supportive, as imagination of the content was prompted. This, in our eyes, would have been an explicit element referring to the third GCL item. Unfortunately, ω dropped to 0.64 when doing so, which is not acceptable anymore. To find out why learners do not rate this item as expected, and not consistently with the other two germane load items, more studies should be conducted to investigate the function of this item. Perhaps the problem with reliability of this item can be found in the nature of prompts and

other germane load enhancing techniques: When learners use them for the first time, they might not see them as helpful and activating in a positive sense. Using a prompt for the first time might result in higher effort in trying to solve the given task, but might—at the same time—be demanding, and does not feel supportive. This might be because of the so-called mathemathantic effect (Clarck 1990): "Instructional treatments unintentionally produced a condition where students were less able to use learning skills or have less access to knowledge in some domain than before entering the experiment" (Clarck 1990, p. 1). The mathemathantic effect usually is applied to training situations, where a given training in a first step reduces performance. Given the prompts in our Study 5 (Technology – clutch), the prompts could interfere with typically used learning strategies, and therefore reduce the use of learning skills that are normally used. Apart from that, they seem to be helpful, because task performance has not been affected negatively, but they are not rated as supportive subjectively.

The conducted confirmatory factor analysis showed a better model fit for Model 2 assuming three inter-related types of cognitive load compared to Model 1 with one overall load-factor. Model fit was nevertheless lower than presented in Klepsch et al. (2017) and not always ideal. This may be because the six studies comprise real learning settings and are thus less artificial but also less clear-cut as the tasks in the original paper. The number of participants did not allow us to conduct a multi-group or multi-level confirmatory factor analysis. The aggregation over all six studies is a second reason for why we could not get a perfect model fit: Combining all studies results in difficulties of comparability, as the six studies vary for example in time on task (e.g. 3 min in Study 2 vs. 40 min in Study 3) and the given learning task (e.g. problem solving in Study 1 vs. remembering facts and relations in Study 4). Nevertheless, the results show that a replication is possible in less artificial learning environments, even if reliability and model fit are not as perfect as under artificial and very controlled circumstances.

Validity

The second criterion of a good measurement instrument is validity. Validity is the degree to which a test measures what it states to be measuring, but what does that mean in terms of experienced cognitive load? As cognitive load is something highly influenced by learners' prerequisites, it is impossible to secure experienced cognitive load to a specific point on a given scale, based on theoretical assumptions. We can state that one version of a learning material should, in contrast to the other version of the same material, result in higher or lower cognitive load for each type of load, but we are unable to operationalize this effect in absolute numbers. Therefore, for our purpose, validity means there exists a relative differentiation between two versions of a learning material (i.e., learners rate their experienced load on the three scales of the CLQ differently for the two versions). Comparing two completely different materials with different contents, length, complexity, etc. to each other, for example, the materials from Study 3 (mathematics) and Study 5 (clutch) of this paper, in our eyes, is not possible and not meaningful. Therefore, sensitivity to differences in learning materials is much more important than securing it to a point on a scale. The additivity theory of cognitive load states that there might not be enough resources left in working memory when too much ICL and ECL is already imposed on the learners' working memory. Thus, one could conclude that the given scale in the CLQ also might be subjectively different for each participant due to interindividual variances in working memory capacity. Also, their ability to chunk information allows them to reduce complex material,

leaving more resources left for ECL or GCL. In summary, we discuss the findings of our six studies regarding the measuring tools' sensitivity towards the measuring tools' ability to differentiate between different types of loads.

For ICL, we found the scale to be sensitive to differences in complexity of the given task (Study 1 and 2). Even if we did not find differences between the flow and overload conditions in both studies, the results showed that participants were able to differentiate between low and high intrinsic load tasks. As the flow condition means that task difficulty is adopted to the persons skill, the flow conditions stay complex but on a level that can be handled, which is reflected in task performance. The overload conditions are even more complex so that the participant can't handle the task, resulting in significantly lower task performances in both studies. In contrast, whenever the number of elements to be learned and their inter-relatedness stayed the same for the experimental groups (Studies 3–6) we found no differences in perceived ICL.

For ECL, we found the scale to be sensitive for design changes (Studies 3 and 4). Unexpectedly, we also found differences in learners' ECL rating when element interactivity has been varied (Study 1). Sweller (2010b) argued that variations in element interactivity might have an impact on all types of load and thus it might not be possible to measure ICL and ECL differentially. The results of the Tetris study seem to provide evidence for this assumption. Unfortunately, faster gameplay seems to also impact ECL. But which aspect of the increased element interactivity causes ICL and which one ECL in our study? We operationalized the increased element interactivity by reducing the time to react, so what does faster gameplay mean with respect to cognitive processing? The blocks come down faster, the preview of he next block is shown shorter, and the keys need to be tapped quicker in order to rotate and move the blocks. Therefore, interaction with the game becomes more complex and perhaps also leads to a higher stress level. The handling of the game, the tapping and moving operations on the keyboard can be seen as extraneous, while the mental operations of filling lines while taking all the elements that interact into account can be seen as intrinsic. However, as both, the mental operation as well as the manual operation has to be aligned, the learners cannot easily distinguish the two sources of load, resulting in increased scores on both scales. To prove this argument, one would have to run a Tetris experiment where learners simply have to mentally rotate the blocks. This would allow disentangling intrinsic from extraneous affordances, even if the design of such a study would be quite a challenge. An additional source of extraneous load might also arise from managing the additional stress level and thus maybe from the suppression of negative emotions or at least an uncomfortable level of arousal. However, such aspects would not be addressed in the questionnaire we used to measure ECL and hence this argument would also need further analyses to be substantiated. Overall, based on studies by Leppink et al. (2013, 2014), the study of Klepsch et al. (2017) and also the knowledge quiz study in this paper we would nevertheless argue that it is possible to distinguish ICL and ECL by psychometric measures. However, when ICL and ECL are intertwined, a differentiated task analysis and maybe also more sophisticated instruments would be necessary to understand the underlying processes and too disentangle intrinsic or extraneous load sources.

Also, the GCL scale was sensitive to differences attempting to induce different levels of GCL (Studies 5 and 6) and could not find any differences when a variation of GCL was not intended (Studies 1–4). However, the validity of the scale is not overall convincing, as we did not find differences in task performance. This would have been crucial from a theoretical point of view, as an increase in GCL should come along with an increase in task performance. For the same reason Leppink et al. (2013) did not accept their measurement of GCL to be valid and useful. The interplay between perceived load and performance

might be more complex and should be analyzed more in-depth: As expected, for Studies 1–4 we found theoretically assumed differences in task performance. Load was lower when reaching higher levels of task performance: In Studies 1–4, lower reported ICL and/or ECL came along with higher task performance. In Studies 5 and 6, we found no significant differences in task performance. From a theoretical point of view, resulting from the definitions of GCL, we would have expected that task performance should be significantly higher when GCL increases. From a practical point of view this is not always true: Learners might invest germane cognitive load into learning and try really hard, but nevertheless fail when they need to retrieve facts and relations from memory and deal with given tasks. Using prompts or the more subtle hint of disfluent text to induce GCL might, as already mentioned, have interfered with their usual learning strategies. So even if they feel to have invested as much GCL as possible, there is no guaranty that they can retrieve their knowledge from memory when needed. Future research could address this complex interplay between the processes that can be triggered by instruction, the perceived germane, extraneous and intrinsic load and the resulting learning performance.

To further substantiate the prognostic validity of the CLQ, it would be interesting to combine it with other differentiated measurement methods of cognitive load during learning. The questionnaire by Leppink et al. (2013) especially could be used to demonstrate concurrent validity. Unfortunately, most of the studies presented in this paper have been conducted before the questionnaire of Leppink et al. (2013) was published. Using the item invented by Paas (1992) would not have been fruitful to use as criterion for concrete validity, because it does not differentiate between ICL, ECL, and GCL. As an example: in Study 6 we would not have been able to show if text legibility is reflected in ECL or GCL.

Strengths and weaknesses of the six studies

All studies in this paper have been designed to explicitly and only vary one aspect of load. In real life learning material this usually is not given. For example, Rogers et al. (2014) used the questionnaire published by Klepsch et al. (2017) in their study on learning piano. They developed a novel media-based piano learning system and compared it to an existing media system and standard sheet notation. Varying the system resulted in significant differences for ICL, ECL, and GCL because motivation and self-regulation during learning had a high impact on cognitive load.

Once again: The six studies are classic experimental studies. Three of them included many more female than male participants, but fortunately, distribution over experimental groups was equal within each study. Two studies included less than 20, but had at least 15 participants in each experimental group. We are aware that the number of participants in each experimental group is not ideal and results in power restrictions. Further research, for example doing a study in a large lecture with many students under real-life conditions, could help overcome this problem and could also provide insights as to whether the questionnaire can be used to evaluate learning material. The low number of participants is especially relevant in the presented confirmatory factor analysis, where the number of participants restricted us to run an overall confirmatory factor analysis and neither nesting nor grouping was possible. As showing the underlying factor structure of the questionnaire was not our primary intention with this paper, a second check to confirm the findings of Klepsch et al. (2017) with a larger sample would be of interest for the whole instructional design community. Covering a variety of learning domains can be seen as positive as well as negative. On the one hand, comparability of

the studies is hardly given because the amount of load between the studies cannot be ranked. On the other hand, the different domains demonstrate that the CLQ can be easily used with different learning materials, environments, and domains. One only must be careful that learners are explicitly instructed which aspect they should rate. Especially in rich learning environments, one person might only be interested in the effectiveness of one presented video, whereas another might be interested in the whole environment. But as stated by Klepsch et al. (2017), the items can easily be adopted to direct the learners to the intended object of investigation.

Especially the studies varying ICL might be questionable as they are not classical learning tasks. First within both studies we did not assess prior knowledge and second the tasks involve no classical learning processes with the goal of schema construction. Nevertheless, both programs used are paradigms in flow research (Keller and Bless 2008; Keller et al. 2011): For both programs Keller et al. (2011; Keller and Bless 2008) could show that participants in the boredom-conditions have skills exceeding the task demands while in the flow-condition they experience a fit between their skills and the task demands and in the overload condition the task demands are perceived as too high compared to one's skills. With reference to load this would mean that the intrinsic load should be rated low for the boredom condition, as it is evaluated with reference to the task affordances in relation to prior knowledge or prior experience (i.e. the skills). Respectively the flow condition should be rated intermediate in ICL and the overload condition as high. These increasing judgments could actually be substantiated in a study of Westphal and Seufert (2016). In the Tetris study, the flow condition was designed with adapting demands to learner's skills. The other two conditions are programmed so skills and demands do not fit together. In the boredom conditions skill exceeds demands, whereas in the overload conditions demands exceed skills. This could be linked to prior knowledge: Even if we did not assess prior knowledge, based on the different playing modes we can assume that prior knowledge, or better prior skills, were higher than needed in the boredom condition, just perfect in the flow condition and too low in the overload condition of the Tetris games. For the quiz-program our argument is similar. First, in this study only participants with at least a diploma from German secondary school qualifying for university admission or matriculation were allowed to participate. Therefore, we can assume similar general knowledge. Based on this, sufficient general knowledge is available, so skills exceed demands in the boredom condition. The flow condition is adapting to the participants general knowledge. The overload condition includes, as is classic for this game, very specific knowledge questions, which should exceed general knowledge of rather fresh high school graduates. They only got right about one third of the questions which is just marginally above guess probability. To sum up, for both studies, even if we did not access something like prior knowledge, we would like to note that the programs are developed to control this objectively by themselves.

Further directions for the use of a differentiated measurement of cognitive load

The presented six studies are more meaningful than the rather artificial learning scenarios used by Klepsch et al. (2017) to develop their questionnaire. But the present studies have still been done in the lab. As a next step, it would be necessary to use the questionnaire in a real-life setting. Therefore, we tend to implement it into a classroom setting within a university setting: for example, students of natural sciences are in need of participation in lab practice, where they need to conduct experiments. Often the experiment is shown by a tutor and has to be repeated by the students. Training the tutors using cognitive apprenticeship

(Collins et al. 1987) might help reduce ECL and foster GCL because students learn what, and especially why, the tutor is doing something while conducting the experiment.

As described by Seufert (2018), it is not only learning the content that causes load on learners' working memory. Also, self-regulatory demands, such as planning, monitoring, and evaluation, can be seen as a source of imposed cognitive load on the learners' working memory. In her paper, Seufert (2018) analyses which self-regulatory activities could cause which type of load. For further analyses of these relations between self-regulation and load the differentiated scale could be helpful. With respect to the present series of studies self-regulatory demands could have also played a role. Asking learners to "try to imagine the learning content as a picture" as done in Study 5, can result in deeper processing of the learning content. In addition, learners might constantly be monitoring whether, or not they reached this goal. This metacognitive monitoring process could be seen as the investment of germane resources, which nevertheless can also be extraneous as they might interfere with the cognitive processes. Thus, learners might report an increased GCL, which might not necessarily lead to higher learning performance. At the very least, cognitive and metacognitive processes compete for the same limited resources.

To sum up, in our six studies, the CLQ to measure different types of imposed cognitive load during learning contributes to the theory of cognitive load and instructional design theories, effects, and principles. Based on our hypotheses and found results, we state that the questionnaire can be rather useful when it is important to differentiate the imposed load on a learner.

Conclusion – what is it good for?

In the last few paragraphs, we argued that the CLQ of Klepsch et al. (2017) is reliable and sensitive to variations in different types of load. But how can a differentiated measurement of cognitive load contribute to the focus of instructional design: optimizing learning. Developing instructional material means (1) deciding on the task, (2) a design, and (3) a method of activating learners' cognitive processes. All three approaches can be fostered by considering theoretical assumptions and studies on instructional design. Usually an increase in learning performance is the goal of instructional design, but this also should be accompanied with the wish to understand the investment of cognitive resources during learning. Therefore, measuring cognitive load during instructional design studies is important to sharpen, confirm, or disprove theoretical assumptions and instructional design effects and principles. Especially measuring cognitive load in studies on instructional design in a differentiated way helps to understand the mechanics of different multimedia principles, as well as aptitude treatment interactions, such as expertise reversal effects. Besides contributing to the understanding of models of learning and instruction and providing evidence for assumptions in studies on multimedia principles, using a differentiated questionnaire to measure aspects of cognitive load can also help to gain insight into effects of learner prerequisites and their impact on individual task performance.

On the one hand, differentially measuring cognitive load helps to understand mechanisms important for learning, as have been proposed by CLT, multimedia principles, or instructional design principles. On the other hand, based on these insights into learning, instructions and materials for learning can be optimized by testing them beforehand.

Understanding theories and effects of learning and instruction

Based on theoretical assumptions of CLT and the Cognitive Theory of Multimedia Learning (CTML; Mayer 2005a), as well as the integrated model of text and picture comprehension (IMTPC; Schnotz 2005), many instructional design principles have been developed and investigated. These principles or effects usually make assumptions on the load that is imposed on the learner, whether they are considered (or not). A systematic testing of these theories and the effects related to them, while also measuring cognitive load in a differentiated way, seems to be fruitful for a deeper understanding of theories and effects of instructional design. At the moment, this is possible with the questionnaires of Leppink et al. (2013, 2014), or Klepsch et al. (2017). For giving practical implications to instructional designers, the traditional view of CLT seems even more helpful than the refined view, which makes using the CLQ by Klepsch et al. (2017) the better option. But further research should, as already mentioned, consider a comparison, or even a combination of both questionnaires. Sweller (2010b), for example stated, that element interactivity is associated with ICL, ECL and GCL. Getting more information on this assumption might not be possible when just asking for the experienced overall cognitive load. In one of our studies, while playing Tetris (Study 1 – see also "Validity"), participants reported significantly different amounts of ECL between conditions, even when the design of the game did not change. It just got faster, and therefore interaction became trickier, and information appeared and disappeared more quickly. Gaining expertise in the game might result in less ECL as one gets used to the fast-paced interaction (like pressing different keys rather fast) with the game. Such an enhanced expertise might also come along with the opportunity of chunking elements. Therefore, expertise might also result in less ICL in higher levels of the game, which brings us to the importance of individual learning processes.

Understand individual learning processes

Individual learning processes and aptitude treatment interactions are important parts of the already mentioned advantages of measuring cognitive load differentially, to confirm or reject theoretical assumptions. For example, measuring cognitive load differentially can be fruitful when comparing experts and novices. Asking only for invested mental effort could result in high load for both groups. Whereas, for example, signaling could ease search processes for novices, and therefore lower ECL, and thus enable them to invest germane resources. For experts' signals could interfere with their prior knowledge and strategies, and therefore result in high ECL. As this example shows, learners' prerequisites are important for understanding effects of instructional design principles. Especially when it comes

to adaption of learning material to learners' prerequisites and prior knowledge, it is important to test the given materials with the target group of learners (e.g., experts/novices, or learners with high/low working memory capacity), and to get feedback on their invested cognitive resources. This might be especially important when fading out assistance in learning. Only through a differentiated measure it is possible to find the correct point to start fading out assistance. Otherwise, one might always have to guess whether, for example, less performance after fading out assistance happens because the new strategy is not yet fluently available, was not understood, or could be applied, but the next step was just too difficult to manage because the last step was not practiced enough.

Getting feedback to optimize instructions and materials for learning

Based on the knowledge about different principles, instructors are able to design effective learning material. But only an evaluation of their content and material can provide insight here and offer concrete feedback. A differentiated measurement of cognitive load in an evaluation of learning material can bring us closer to our goal of designing an effective learning experience by finding problems and stumbling blocks. In a rich learning environment (e.g., an online learning environment), measuring load could also be fruitful in combination with measurement methods of user experience. Rogers et al. (2014), as already mentioned, investigated cognitive load in a computerized learning environment for playing the piano. They developed a system with a projected roll notation of notes in order to learn how to play piano. While applying some well-known instructional design principles in smaller and specialized parts of their program (split-attention effect, testing effect, segmenting principle, signaling etc.), there was no evidence offered on how participants would react to the whole system. By combining a user experience questionnaire with the CLQ, they gained insight into cognitive resources devoted to learning during the system usage, as well as insights into the user experience of their whole learning environment.

In summary, we argue that a differentiated measurement of cognitive load in instructional design research is necessary. Based on the question of instructional designers presented at the beginning, the differentiated measurement of cognitive load, on the one hand, helps researchers to understand theories and design principles in more depth, especially when it comes to the role of learners' prerequisites, such as prior knowledge, working memory capacity, or spatial abilities. On the other hand, it helps instructors and designers to evaluate parts of their learning material and content (e.g., one instructional video), or – in combination with other methods – in evaluating rich learning material as a whole (e.g., a complex online learning environment).

Acknowledgements We like to thank our former students involved in one of the six studies. They contributed to this research as part of their curriculum, in form of a bachelor or master thesis, or in exertion of their position as student research assistants.

Author contributions MK and TS contributed to the conception and design of the studies. MK developed the used material and questionnaires, and led data collection for studies 1–4. TS developed the used material and questionnaires, and led data collection for studies 5 and 6. MK analyzed and interpreted the data. MK drafted the work, which was revised critically by TS. All authors provided approval of the final submitted version of the manuscript and agreed to be accountable for all aspects of the work by ensuring that questions related to the accuracy or integrity of any part of the work was appropriately investigated and resolved.

Funding This work was supported by the Federal Ministry of Education and Research Germany within the program "Aufstieg durch Bildung: Offene Hochschulen" as part of the project "Effizient Interaktiv Studieren II (EffIS-II)" (grant number 16OH22032). Open Access funding provided by Projekt DEAL.

Compliance with ethical standards

Ethical approval All six studies were exempt from an ethic committee approval due to the recommendations of the German Research Association: No subject was in risk of physical or emotional pressure, we fully informed all subjects in all studies about the goals and processes of the study they participated in, and none of the subjects were patients, minors, or persons with disabilities. In all studies, participation was voluntary, and all subjects signed a written informed consent and were aware that they had the chance to withdraw their data at any point of the study.

Appendix 1

See Table 10.

Table 10 CLQ from Klepsch et al. (2017)

Type of load	Item – German	Item – English
ICL	Bei der Aufgabe musste man viele Dinge gleichzeitig im Kopf bearbeiten	For this task, many things needed to be kept in mind simultaneously
ICL	Diese Aufgabe war sehr komplex	This task was very complex
GCL	Ich habe mich angestrengt, mir nicht nur einzelne Dinge zu merken, sondern auch den Gesamtzusammenhang zu verstehen	I made an effort, not only to understand several details, but to understand the overall context
GCL	Es ging mir beim Bearbeiten der Lerneinheit darum, alles richtig zu verstehen	My point while dealing with the task was to understand everything correct
GCL*	Die Lerneinheit enthielt Elemente, die mich unterstützten, den Lernstoff besser zu verstehen	The learning task consisted of elements supporting my comprehension of the task
ECL	Bei dieser Aufgabe ist es mühsam, die wichtigsten Informationen zu erkennen	During this task, it was exhausting to find the important information
ECL	Die Darstellung bei dieser Aufgabe ist ungünstig, um wirklich etwas zu lernen	The design of this task was very inconvenient for learning
ECL	Bei dieser Aufgabe ist es schwer, die zentralen Inhalte miteinander in Verbindung zu bringen	During this task, it was difficult to recognize and link the crucial information

Note. ICL intrinsic cognitive load; *ECL* extraneous cognitive load; *GCL* germane cognitive load. *Item only useful if GCL is varied on purpose in a given learning material (e.g. through prompts)

Appendix 2

See Table 11.

Table 11 Correlations between control and dependent variables for all 6 studies

	ICL	ECL	GCL	Task performance
Play Tetris (Study 1)				
Age	.19	.10	.24	− .18
Answer knowledge question (Study 2)				
Age	.08	− .11	− .07	.25
Solve mathematical problems (Study 3)				
Age	.20	.14	.15	− .38*
Prior knowledge	− .26	− .20	− .138	.30
Learn about the kidney (Study 4)				
Age	.03	− .42**	.09	− .05
Prior knowledge	− .12	− .17	.29*	.43**
Learn about the clutch (Study 5)				
Age	.07	− .12	.14	− .06
Prior knowledge	− .51**	− .19	.16	.40**
FAM – anxiety of failure	.09	− .31*	.02	.05
FAM – probability of success	.07	− .18	− .09	.04
FAM – interest	− .05	− .43*	.27*	.40*
FAM – challenge	.17	− .30*	.20	.42**
Time on task	.31*	.01	.26*	.18
Earth's time zones (Study 6)				
Age	.16	− .02	− .18	− .21
Prior knowledge	− .50*	− .34*	− .35*	.49**
FAM – anxiety of failure	.21	.19	.06	− .31
FAM – probability of success	− .28	− .43**	.17	.19
FAM – interest	.13	− .10	.23	.11
FAM – challenge	.21	.10	.33	− .19

Note $^{*}p < .05$; $^{**}p < .01$

References

Alter, A. L., & Oppenheimer, D. M. (2009). Uniting the tribes of fluency to form a metacognitive nation. *Personality and Social Psychology Review, 13*(3), 219–235.

Artino, A. R., Jr. (2008). Cognitive load theory and the role of learner experience: An abbreviated review for educational practitioners. *AACE Journal, 16*(4), 425–439.

Atkinson, R. K., Derry, S. J., Renkl, A., & Wortham, D. (2000). Learning from examples: Instructional principles from the worked examples research. *Review of Educational Research, 70*(2), 181–214.

Ayres, P. (2006). Impact of reducing intrinsic cognitive load on learning in a mathematical domain. *Applied Cognitive Psychology, 20*(3), 287–298.

Ayres, P., & Sweller, J. (2005). The split-attention principle in multimedia learning. In R. E. Mayer (Ed.), *The Cambridge handbook of multimedia learning* (pp. 135–146). Cambridge, New York: Cambridge University Press.

Bannert, M. (2009). Promoting self-regulated learning through prompts. *Zeitschrift für Pädagogische Psychologie, 23*(2), 139–145.

Brünken, R., Plass, J. L., & Leutner, D. (2003). Direct measurement of cognitive load in multimedia learning. *Educational Psychologist, 38*(1), 53–61.

Brünken, R., Seufert, T., & Paas, F. G. W. C. (2010). Measuring cognitive load. In J. L. Plass, R. Moreno, & R. Brünken (Eds.), *Cognitive load theory* (pp. 181–202). Cambridge, New York: Cambridge University Press.

Chandler, P., & Sweller, J. (1991). Cognitive load theory and the format of instruction. *Cognition & Instruction, 8*(4), 293–332.

Clarck, R.E. (1990). When teaching kills learning. research on mathemathantics. In H. Mandl, E. de Corte, N. Bennett, & H.F. Friedrich (Eds.), *Analysis of complex skills and complex knowledge domains: Learning and instruction: European research in an international context: Volume II*: Selection of papers from the Second European Conference for Research on Learning and Instruction held in Tübingen, West Germany, in September 1987 (pp. 1–22). Oxford: Pergamon Press.

Collins, A., Brown, J.S. & Newman, S.E. (1987). *Cognitive apprenticeship. teaching the craft of reading, writing, and mathematics.* Technical Report.

Debue, N., & van de Leemput, C. (2014). What does germane load mean? An empirical contribution to the cognitive load theory. *Frontiers in Psychology, 5*, 1099.

Eitel, A., Kühl, T., Scheiter, K., & Gerjets, P. (2014). Disfluency meets cognitive load in multimedia learning: Does harder-to-read mean better-to-understand? *Applied Cognitive Psychology, 28*(4), 488–501.

Fletcher, J. D., & Tobias, S. (2005). The multimedia principle. In R. E. Mayer (Ed.), *The Cambridge handbook of multimedia learning* (pp. 117–134). Cambridge, New York: Cambridge University Press.

Jumbo Spiele®,. (2000). *Wer wird Millionär? [Who wants to be a millionaire?].* Herscheid (DE): Jumbo Spiele GmbH.

Kalyuga, S. (2011). Cognitive load theory: How many types of load does it really need? *Educational Psychology Review, 23*(1), 1–19.

Kalyuga, S., Chandler, P., & Sweller, J. (1998). Levels of expertise and instructional design. *Human Factors: The Journal of the Human Factors and Ergonomics Society, 40*(1), 1–17.

Keller, J., & Bless, H. (2008). Flow and regulatory compatibility: An experimental approach to the flow model of intrinsic motivation. *Personality & Social Psychology Bulletin, 34*(2), 196–209.

Keller, J., Bless, H., Blomann, F., & Kleinböhl, D. (2011). Physiological aspects of flow experiences: Skills-demand-compatibility effects on heart rate variability and salivary cortisol. *Journal of Experimental Social Psychology, 47*(4), 849–852.

Klepsch, M., Schmitz, F., & Seufert, T. (2017). Development and validation of two instruments measuring intrinsic, extraneous, and germane cognitive load. *Frontiers in Psychology, 8*, 1997.

Kühl, T., & Eitel, A. (2016). Effects of disfluency on cognitive and metacognitive processes and outcomes. *Metacognition and Learning, 11*(1), 1–13.

Leppink, J., Paas, F. G. W. C., van der Vleuten, C. P. M., van Gog, T., & van Merriënboer, J. J. G. (2013). Development of an instrument for measuring different types of cognitive load. *Behavior research methods, 45*(4), 1058–1072.

Leppink, J., Paas, F. G. W. C., van Gog, T., van der Vleuten, C. P. M., & van Merriënboer, J. J. G. (2014). Effects of pairs of problems and examples on task performance and different types of cognitive load. *Learning and Instruction, 30*, 32–42.

Mayer, R. E. (2005a). Cognitive theory of multimedia learning. In R. E. Mayer (Ed.), *The Cambridge handbook of multimedia learning* (pp. 31–48). Cambridge, New York: Cambridge University Press.

Mayer, R. E. (Ed.). (2005b). *The Cambridge handbook of multimedia learning*. Cambridge, New York: Cambridge University Press.

Mayer, R. E. (2009). *Multimedia learning* (2nd ed.). Cambridge: Cambridge University Press.
Moreno, R., & Mayer, R. E. (1999). Cognitive principles of multimedia learning: The role of modality and contiguity. *Journal of Educational Psychology, 91*(2), 358–368.
Moreno, R., & Park, B. (2010). Cognitive load theory: Historical development and relation to other theories. In J. L. Plass, R. Moreno, & R. Brünken (Eds.), *Cognitive load theory* (pp. 9–28). Cambridge, New York: Cambridge University Press.
Paas, F. G. W. C. (1992). Training strategies for attaining transfer of problem-solving skill in statistics: A cognitive-load approach. *Journal of Educational Psychology, 84*(4), 429–434.
Plass, J. L., Moreno, R., & Brünken, R. (Eds.). (2010). *Cognitive load theory*. Cambridge, New York: Cambridge University Press.
Pollock, E., Chandler, P., & Sweller, J. (2002). Assimilating complex information. *Learning and Instruction, 12*(1), 61–86.
Renkl, A. (2005). The worked-out examples principle in multimedia learning. In R. E. Mayer (Ed.), *The Cambridge handbook of multimedia learning* (pp. 229–246). Cambridge, New York: Cambridge University Press.
Renkl, A., & Atkinson, R. K. (2010). Learning from worked-out examples and problem solving. In J. L. Plass, R. Moreno, & R. Brünken (Eds.), *Cognitive load theory* (pp. 91–108). Cambridge, New York: Cambridge University Press.
Rheinberg, F., Vollmeyer, R., & Burns, B. D. (2001). FAM: Ein Fragebogen zur Erfassung aktuller Motivation in Lern- und Leistungssituationen. *Diagnostica, 47*(2), 57–66.
Rogers, K., Röhlig, A., Weing, M., Gugenheimer, J., Könings, B. & Klepsch, M., et al. (2014). P.I.A.N.O. In R. Dachselt, N. Graham, K. Hornbæk, & M. Nacenta (Eds.), *ITS '14 Proceedings of the Ninth ACM International Conference on Interactive Tabletops and Surfaces* (pp. 149–158).
Schnotz, W. (2005). An integrated model of text and picture comprehension. In R. E. Mayer (Ed.), *The Cambridge handbook of multimedia learning* (pp. 49–70). Cambridge, New York: Cambridge University Press.
Schnotz, W., & Bannert, M. (1999). Einflüsse der Visualisierungsform auf die Konstruktion mentaler Modelle beim Text- und Bildverstehen. *Experimental Psychology, 46*(3), 217–236.
Seufert, T. (2018). The interplay between self-regulation in learning and cognitive load. *Educational Research Review, 24*, 116–129.
Seufert, T., Wagner, F., & Westphal, J. (2017). The effects of different levels of disfluency on learning outcomes and cognitive load. *Instructional Science, 45*(2), 221–238.
Sweller, J. (1994). Cognitive load theory, learning difficulty, and instructional design. *Learning and Instruction, 4*(4), 295–312.
Sweller, J. (2010a). Cognitive load theory: Recent theoretical advances. In J. L. Plass, R. Moreno, & R. Brünken (Eds.), *Cognitive load theory* (pp. 29–47). Cambridge, New York: Cambridge University Press.
Sweller, J. (2010b). Element interactivity and intrinsic, extraneous, and germane cognitive load. *Educational Psychology Review, 22*(2), 123–138.
Sweller, J., & Chandler, P. (1994). Why some material is difficult to learn. *Cognition & Instruction, 12*(3), 185.
Sweller, J., van Merriënboer, J. J. G., & Paas, F. G. W. C. (1998). Cognitive architecture and instructional design. *Educational Psychology Review, 10*(3), 251–296.
Sweller, J., Ayres, P., & Kalyuga, S. (2011). *Cognitive load theory*. New York: Springer.
van Merriënboer, J. J. G., & Ayres, P. (2005). Research on cognitive load theory and its design implications for E-learning. *Educational Technology Research and Development, 53*(3), 5–13.

Klepsch, M., Kempter, I. & Seufert, T. (2013). Interaction of Worked Examples and Prompts: Impact on Performance and Cognitive Load. In European Association for Research on Learning and Instruction (Hrsg.), *15th Biennial EARLI Conference "Responsible Teaching and Sustainable Learning"*, München (DE).

Abrufbar unter: https://www.earli.org/book-abstracts

European Association for Research on Learning and Instruction (EARLI) Conference 2013 –
Accepted Submission

Interaction of Worked Examples and Prompts: Impact on Performance and Cognitive Load

Melina Klepsch
Ulm University, Institute of Psychology and Education
Albert-Einstein-Allee 47, 89081 Ulm, Germany
melina.klepsch@uni-ulm.de

Isabel Kempter
Ulm University, Institute of Psychology and Education
Albert-Einstein-Allee 47, 89081 Ulm, Germany
isabel.kempter@uni-ulm.de

Tina Seufert
Ulm University, Institute of Psychology and Education
Albert-Einstein-Allee 47, 89081 Ulm, Germany
tina.seufert@uni-ulm.de

Abstract.
To investigate the interaction between worked examples and self-explanation-prompts we conducted an experimental study with 73 learners. In the present study we investigated the impact on learning performance and cognitive load. We assume that worked examples and prompts will cause better learning performance, worked examples will reduce extraneous cognitive load, and prompts will foster germane cognitive load. In a mathematical domain after learning, the subjects had to perform analog tasks and transfer tasks. We could show the advantage of worked examples and their reducing impact on extraneous load. For prompts we could not show a significant impact on learning outcome or higher germane load.

Keywords: Comprehension of Text and Graphics, Instructional Design
SIG: 2 – Comprehension of Text and Graphics
Domain: Learning and Cognitive Science
Scheduling categories: Comprehension of Text and Graphics

Introduction and Theoretical Background

As theoretical framework for the presented study we used cognitive load theory (CLT; Chandler & Sweller, 1991) and the adaptive control of thought - rational theory (ACT-R; Anderson, Fincham, & Douglass, 1997). Both theories explain the effectiveness of worked examples and self-explanation-prompts, which have been manipulated in the present study.

ACT-R theory says acquiring new skills takes place in four steps: At the beginning learners solve problems by analogies, then abstract rules and schemata are used. Afterwards learners are able to solve known problems automatically. In the end they are able to solve all kinds of problems. According to Atkinson et al. (Atkinson, Derry, Renkl, & Wortham, 2000) worked examples are useful in a first phase of skill acquisition. Studies based on CLT revealed that worked examples result in better performance and less extraneous load compared to problem solving (e.g. Atkinson et al., 2000). Moreover, the effectiveness of worked examples seems to be influenced by self-explanations (Chi, Bassok, Lewis, Reimann, & Glaser, 1989). However, self-explanations are rarely generated when not explicitly prompted (Renkl, 1997). Therefore, we initiated self-explanations by presenting prompts. Prompts can be seen as requests requiring students to process the learning material in a specific way (Clark, Nguyen, & Sweller, 2006). For prompts we assume assistance by activating prior knowledge.

In our study we investigated the interaction between worked examples and self-explanation-prompts. For learning outcome we assume that the combination of both approaches results in better

performance compared to learning material only consisting of worked-examples or self-explanation-prompts. Students without assistance should perform worst.
For cognitive load we hypothesize that worked examples will reduce extraneous cognitive load (ECL), whereas prompts increase germane cognitive load (GCL). We assume no changes in intrinsic cognitive load (ICL).

Method

In our study 73 students took part and were randomly assigned to one of four groups: (a) problem solving without prompts (n=20), or (b) with prompts (n=15), and (c) worked examples without prompts (n=20), or (d) with prompts (n=19). Each learner had to deal with two mathematical learning materials (extreme value and Taylor polynomial tasks). Performance was measured by a post test for each domain, containing analog tasks, similar to the learning material, and complex transfer tasks. Cognitive load was measured after learning and during the posttest by a questionnaire which differentiated ICL (Cronbach's α=.92), ECL (Cronbach's α=.8) and GCL (Cronbach's α=.86).

Results

Learning performance

For analog tasks we found a main effect for worked examples ($F(1,69)=11.48$, $p_1<.001$, $\eta^2=0.14$). For prompts no main effect could be found ($F(1,69)=1.15$, n.s.) and also no interaction ($F(1,69)=1.15$, n.s.). See figure 1. We couldn't find any effects for transfer tasks (Fs< 1, n.s).

Cognitive load during the learning phase

Concerning ICL we couldn't find a main effect for worked examples nor for prompts (Fs< 1, n.s.), and also no interaction ($F(1,69)=1.39$, n.s.). For ECL we found a main effect of worked examples ($F(1,69)=12.47$, $p1<.001$, $\eta^2=0.15$), but no main effect for prompts, and also no interaction (Fs< 1, n.s). For GCL we found no main effect of worked examples (F< 1, n.s) but by trend a main effect for prompts ($F(1,69)=1.84$, $p_1=.09$, $\eta^2=0.03$). An interaction couldn't be found (F< 1, n.s). See figure 2.

Cognitive load during the post tests

For ICL no effects were found. (Fs< 1, n.s). For ECL we found a main effect for worked examples ($F(1,69)=14.23$, $p_1<.001$, $\eta^2=0.17$), but no main effect for prompts, and no interaction (Fs< 1, n.s). For GCL we found no effect of worked examples ($F(1,69)=1.76$, n.s.), but by trend a main effect for prompts ($F(1,69)=2.01$, $p_1=.08$, $\eta^2=0.03$). We couldn't show an interaction ($F(1,69)=1.05$, n.s.). See figure 3.

Summary and Discussion

Our study investigated the combination of worked examples and self-explanation-prompts. As assumed, worked examples increased learning performance for analog tasks. Perhaps due to task complexity, we couldn't find the effect for transfer tasks. ACT-R theory explains this finding: Only practice is able to bring you on a level which stands for the ability to solve all kinds of problems (Anderson et al., 1997). In our study students only had to deal with an introducing text, and two examples (shown as mean-ends problems or worked examples), followed by two analog tasks, and one transfer task. This might be a too short time to reach a level where transfer tasks can be solved adequately.
As predicted, worked examples result always in a reduction of ECL.

For GCL we found inconsistent results: In contrast to our prediction, after the learning phase learners with prompts reported by trend less GCL than learners without prompts. After the posttest learners with prompts reported by trend more GCL than learners without prompts in the learning phase. A possible explanation may be found in the time limit for the learning material. Perhaps time was too short, so learners were not able to deeply deal with the presented prompts.

As complexity of the learning content didn't vary, we postulated and could confirm that ICL isn't varying between groups.

References

Anderson, J. R., Fincham, J. M., & Douglass, S. (1997). The role of examples and rules in the acquisition of a cognitive skill. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *23*(4), 932–945.

Atkinson, R. K., Derry, S. J., Renkl, A., & Wortham, D. (2000). Learning from Examples: Instructional Principles from the Worked Examples Research. *Review of Educational Research*, *70*(2), 181–214.

Chandler, P., & Sweller, J. (1991). Cognitive Load Theory and the Format of Instruction. *Cognition & Instruction*, *8*(4), 293–332.

Chi, M. T., Bassok, M., Lewis, M. W., Reimann, P., & Glaser, R. (1989). Self-Explanations: How Students Study and Use Examples in Learning to Solve Problems. *Cognitive Science*, *13*(2), 145–182.

Clark, R. C., Nguyen, F., & Sweller, J. (2006). *Efficiency in learning: Evidence-based guidelines to manage cognitive load*. San Francisco: Pfeiffer.

Renkl, A. (1997). Learning from Worked-Out Examples: A Study on Individual Differences. *Cognitive Science*, *21*(1), 1–29.

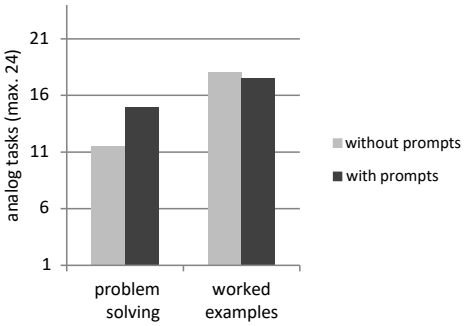

Figure 1: points for analog tasks

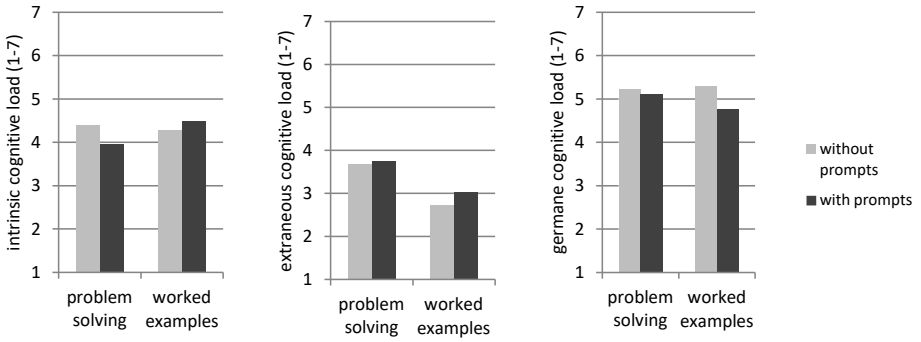

Figure 2: cognitive load during learning.

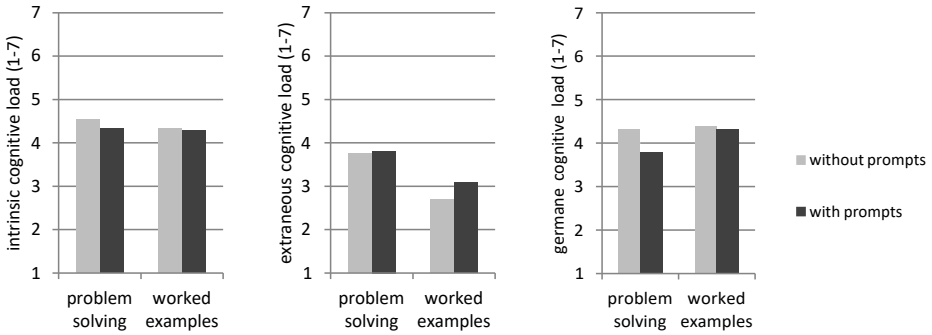

Figure 3: cognitive load during the posttest.

Klepsch, M., Westphal, J. & Seufert, T. (2013). Effekte von integriertem bzw. separiertem Text-Bild-Material in Abhängigkeit von serialistischem oder holistischem Lernstil., *14. Fachgruppentagung Pädagogische Psychologie*. Hildesheim (DE).

Ursprünglich Abrufbar unter: http://www.paeps-hildesheim-2013.de [nicht mehr erreichbar]

Gemeinsam verschieden. 14. Fachgruppentagung Pädagogische Psychologie 2013
Akzeptierte Einreichung

Effekte von integriertem bzw. separiertem Text-Bild-Material in Abhängigkeit von serialistischem oder holistischem Lernstil

Klepsch, Melina; Westphal, Julia; Seufert, Tina

Bilder können den Lernprozess unterstützen, wenn sie mit einem Text kombiniert werden, der entweder neben dem Bild oder im Bild platziert sein kann. Die separierte Darstellung führt laut zahlreicher Studien zu einem split attention effect, der sowohl die kognitive Belastung erhöht als auch den Lernerfolg behindert, da die Lernenden ihre Aufmerksamkeit aufteilen müssen, was Such- und Orientierungsprozesse erfordert. In einem 2x2-faktoriellen Design (n=51) wurde neben dem Design-Faktor (split vs. integriert) der Lernstil serialist vs. holist (Pask, 1969) als Faktor untersucht (Lernerfolg und kognitive Belastung als AVn). Probanden mit split-source Material waren signifikant höher belastet und zeigten geringeren Lernerfolg als mit integriertem Material. In Bezug auf den Lernstil zeigte sich, dass Serialisten generell bessere Lernleistungen und geringere Belastung zeigten, die Holisten diesen Nachteil aber kompensieren konnten, wenn sie mit integriertem Lernmaterial arbeiten konnten.

Seufert, T., Klepsch, M. & Westphal, J. (2017). Adjusting Task Difficulty to Control Learner's Intrinsic Cognitive Load. In European Association for Research on Learning and Instruction (Hrsg.), *EARLI 2017 Ëducation in the crossroads of economy and politics - Role of research in the advancement of public good"*, Tempere (FIN).

Abrufbar unter: https://www.earli.org/book-abstracts

European Association for Research on Learning and Instruction (EARLI) Conference 2017 – Accepted Submission

Adjusting Task Difficulty to Control Learner's Intrinsic Cognitive Load

Melina Klepsch
Ulm University, Institute of Psychology and Education
Albert-Einstein-Allee 47, 89081 Ulm, Germany
melina.pichler@uni-ulm.de

Julia Westphal
Ulm University, Institute of Psychology and Education
Albert-Einstein-Allee 47, 89081 Ulm, Germany
julia.westphal@uni-ulm.de

Tina Seufert
Ulm University, Institute of Psychology and Education
Albert-Einstein-Allee 47, 89081 Ulm, Germany
tina.seufert@uni-ulm.de

Abstract

A lot of studies exist on varying extraneous cognitive load and thereby testing different multimedia design principles. Studies on germane and especially on intrinsic cognitive load variations, are less popular. Most intrinsic load studies use a categorization of learners into novices and experts to vary intrinsic load, very often in relation to interactions with different treatment effects. In all these cases intrinsic load has not been levelled to learners' skills individually. This may be due to the problem of producing treatment material, that adapts to the learner's skills and prior knowledge. Therefore, in three experimental studies with learning and testing material originally coming from flow-research, we tested intrinsic load variations by adapting the demands of the task to the learner's skills and prior knowledge. In three studies different domains (mathematics, gameplay, general knowledge) are addressed and the learning and testing material adjusts the level of difficulty to the participant's skill level. Hence, three experimental groups are generated: (1) boredom (skills exceeding demands), (2) fit (skills matching demands), and (3) overload (demands exceed skills). Preliminary results with n=58 students and the mathematical learning material show promising results that automatically adjusting the level of difficulty to the participant's skill level results in the expected level of intrinsic cognitive load, measured by a subjective differentiated questionnaire of cognitive load.

Keywords: Experimental Studies, Assessment methods and tools, Instructional Design, Cognitive Skills, Comprehension of text and graphics
Domain: Instructional Design
SIG: SIG 02 - Comprehension of Text and Graphics

Introduction and Theoretical Background

Cognitive Load Theory (CLT; Chandler & Sweller, 1991) assumes that working memory is limited in capacity and deals with the three conceptual parts of cognitive load (Sweller, van Merriënboer & Paas, 1998): (a) Intrinsic cognitive load (ICL), results from complexity and element interactivity of the task that must be performed, and the degree of prior knowledge and learner's expertise (Schnotz & Kürschner, 2007). (b) Extraneous cognitive load (ECL) occurs due to the design of the learning environment. (c) Germane cognitive load (GCL) arises from the learner's activities in understanding the learning material and building schemata. Hence, learning material should strive for appropriate ICL, low ECL, and high GCL.

Research often concentrates on ECL variations and testing different multimedia principles that should help to optimize ECL level. Design aspects are highly relevant on the one hand and on the other hand they are easy to manipulate. Variations of ICL are harder to implement, as ICL mainly results from learner's prior knowledge and expertise. Besides some studies that use segmentation to manipulate ICL, most experimental studies "vary" ICL by classifying learners as novices or experts by using prior knowledge tests. Most of these studies use these categories to detect interactions with specific treatment or design effects. In general, this categorization is a rather good possibility to get learners whose ICL should be low (experts) respectively high (novices) during learning. Problem of this method is, that this only results in a categorical disjunction of participant and does not allow for more individual approaches where learners skills are aligned to the tasks demands.

Hence, in our study we liked to give adjusting of task difficulty a try. The idea is that based on learner's performance tasks get automatically easier or harder. Through this procedure ICL can be adapted to the learner's zone of proximal development. This idea results from flow theory (Csikszentmihaly, 1990), which says that individuals function best if a fit of ability and task difficulty exists. To induce flow Keller and Colleagues (Keller & Bless, 2008; Keller, Bless, Blomann, & Kleinböhl, 2011, Keller & Landhäußer, 2011) invented different computerized programs which adapt task difficulty to learner's ability. The programs generate three different load levels and therefore three experimental conditions: (1) boredom (skills exceeding demands), (2) fit (skills matching demands), and (3) overload (demands exceed skills). In our study the question was whether an active manipulation of intrinsic cognitive load based on the task difficulty (a learner can handle) is reflected in the answers of a subjective differentiated measurement of CLT, especially of ICL.

Therefore, we assume an ascending order of reported ICL for the different groups:

H1: ICL of boredom group < ICL of fit group < ICL of overload group

As Design is not changing, we assume no change of ECL over all groups:

H2: ECL of boredom group = ECL of fit group = ECL of overload group

As we address learner's zone of proximal development in the fit group, we expect GCL to be highest when participants are in the fit group:

H3: (GCL of boredom group = GCL of overload group) < GCL of fit group

Method

In the experimental study three computer programs originally designed for flow research, are used to vary ICL. The tasks are (1) a mathematical test (Keller & Landhäußer, 2011), (2) a Tetris game (Keller & Bless, 2008) and (3) a knowledge task based on the German TV show "Who wants to be a millionaire?" (Jumbo Spiele®, 2000; Keller et. al 2011).

In each version three conditions were tested (boredom, fit, overload). Demographic data (age, sex) was assessed, and after the task each participant filled out a subjective differentiated questionnaire on cognitive load (Klepsch & Seufert, 2012), and rated their motivation.

Until now, we can present first analyses with the math program where two conditions have been realized: (1) boredom and (2) fit. The 58 participants were randomly assigned to one of the two conditions (boredom: n=29; fit: n=29).

Results

Preliminary results (Fig. 1) show, according to hypothesis 1, that participants in the boredom group reported significant lower ICL than participants in the fit group ($t(56)=10.62$, $p<.001$, $d=2.79$). We could, according to hypothesis 2, not find a difference in ECL ($t(56)=0.86$, n.s.). Hypothesis 3, could not be confirmed, ($t(56)=0.75$, n.s.).

Summary and Discussion

According to the promising preliminary results we are encouraged to find similar results also for the other versions of the flow programs following the same approach (Tetris, "Who wants to be a millionaire"). We assume these tests, which are adaptive to learner's skills, might be a good way to validate differentiated measurement methods of cognitive load. Adaptive learning material which is fitting to learner's skills, bore someone, or overloads one might also be a good starting point in (re-)testing multimedia principles. With such material that can be adopted to fit a learner's skills variance in prior knowledge according to the to be tested learning material could be minimized within the material.

References

Csikszentmihaly, M. (1990). *Flow*. New York: Harper & Row.

Chandler, P., & Sweller, J. (1991). Cognitive Load Theory and the Format of Instruction. *Cognition and Instruction, 8*, 293-332.

Jumbo Spiele® (2000). *Wer wird Millionär?* Germany: Herscheid.

Keller, J., & Bless, H. (2008). Flow and regulatory compatibility: An experimental test of the flow model of intrinsic motivation. *Personality and Social Psychology Bulletin, 34*, 196−209

Keller, J., Bless, H., Blomann, F., & Kleinböhl, D. (2011). Physiological aspects of flow experiences: Skills-demand-compatibility effects on. *Journal of Experimental Social Psychology, 47*, 849-852.

heart rate variability and salivary cortisol

Keller, J., & Landhäußer, A. (2011). Im Flow sein: Experimentelle Analysen des Zustands optimaler Beanspruchung. *Psychologische Rundschau, 62*, 213-220.

Klepsch, M., & Seufert, T. (2012). Subjective differentiated measurement of cognitive load. Paper presented at the *5th International Cognitive Load Theory Conference*, Tallahassee (USA).

Schnotz, W. & Kürschner, C. (2007). A reconsideration of cognitive load theory. *Educational Psychology Review, 19*, 469-508.

Sweller, J., van Merriënboer, J., & Paas, F. (1998). Cognitive Architecture and Instructional Design. *Educational Psychology Review*, 10(3), 251–296.

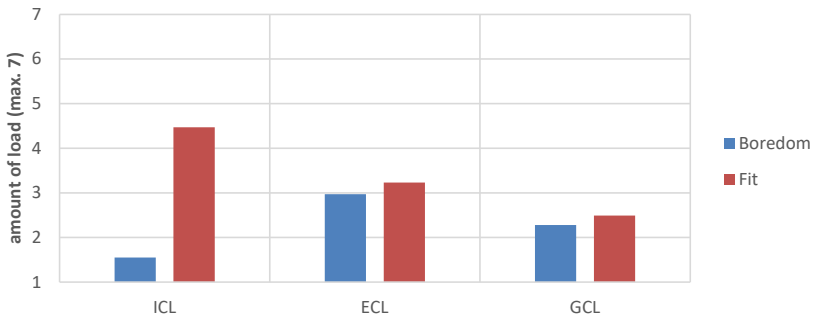

Fig.1: preliminary results for self-reported ICL, ECL and GCL for both groups (boredom, fit)

Rogers, K., Röhlig, A., Weing, M., Gugenheimer, J., Könings, B., Klepsch, M., Schaub, F., Rukzio, E., Seufert, T. & Weber, M. (2014). P.I.A.N.O: Faster Piano Learning with Interactive Projection. In R. Dachselt, N. Graham, K. Hornbæk & M. Nacenta (Hrsg.), *Ninth ACM International Conference on Interactive Tabletops and Surfaces*, Dresden (DE). doi: 10.1145/2669485.2669514

Abrufbar unter: https://doi.org/10.1145/2669485.2669514

P.I.A.N.O.: Faster Piano Learning with Interactive Projection

Katja Rogers[1], Amrei Röhlig[1], Matthias Weing[1], Jan Gugenheimer[1], Bastian Könings[1], Melina Klepsch[2], Florian Schaub[3], Enrico Rukzio[1], Tina Seufert[2], Michael Weber[1]
[1]Institute of Media Informatics, [2]Institute of Psychology and Education
Ulm University, Germany, [firstname.lastname]@uni-ulm.de
[3]School of Computer Science, Carnegie Mellon University, Pittsburgh, fschaub@cs.cmu.edu

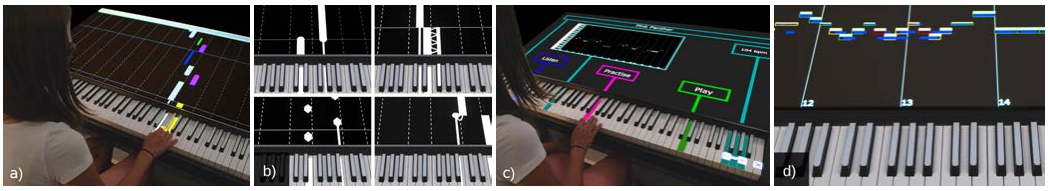

Figure 1. In P.I.A.N.O., (a) music notation is projected onto the piano to facilitate direct mapping of notes to respective piano keys. Correct fingering is supported with color highlights. (b) The basic piano roll notation is extended to support articulations (legato, staccato) and note ornaments (trill, grace notes). (c) The learning process is supported by three learning modes (*listen*, *practice*, *play*). (d) After *practicing* a song, the *play* mode provides detailed feedback on the achieved skill level.

ABSTRACT
Learning to play the piano is a prolonged challenge for novices. It requires them to learn sheet music notation and its mapping to respective piano keys, together with articulation details. Smooth playing further requires correct finger postures. The result is a slow learning progress, often causing frustration and strain. To overcome these issues, we propose P.I.A.N.O., a piano learning system with interactive projection that facilitates a fast learning process. Note information in form of an enhanced piano roll notation is directly projected onto the instrument and allows mapping of notes to piano keys without prior sight-reading skills. Three learning modes support the natural learning process with live feedback and performance evaluation. We report the results of two user studies, which show that P.I.A.N.O. supports faster learning, requires significantly less cognitive load, provides better user experience, and increases perceived musical quality compared to sheet music notation and non-projected piano roll notation.

Author Keywords
Piano; music; instrument learning; interactive projection; piano roll notation; muscial expression; CAMIT.

ACM Classification Keywords
H.5.5 Information Interfaces and Presentation: Sound and Music Computing; K.3.1 Computers and Education: Computer Uses in Education—*Computer-assisted instruction*.

ITS 2014, November 16–19, 2014, Dresden, Germany.

http://dx.doi.org/10.1145/2669485.2669514

INTRODUCTION
Hallam [13] has shown that children and young people benefit from the positive effects of music-making on personal and social development. Learning to play piano as an adult is often motivated by self-actualization and enjoyment [18]. Active music-making has also been found to enhance the health and well-being of elderly people, and can even contribute to recovery from depression [5]. However, those positive effects will only occur when playing is enjoyable and rewarded [13].

For novices, learning to play piano is a prolonged challenge. One reason for this is that conventional sheet music notation (e.g., the western notation based on a five-line staff) does not support the mapping of notes to piano keys very well [32, 23]. First, reading sheet music, called *sight-reading*, requires a vertical to horizontal mapping of notes' pitch information to the respective piano keys. Second, it requires learners to map complex note symbols and articulation marks to the notes' duration and their individual expression. While trained pianists process this mapping automatically, it is a burdensome challenge for novices [28]. Furthermore, in order to play accurately, correct *fingering* is required, i.e., which fingers to use for specific notes and note sequences [27]. These burdens often result in frustration and high dropout rates from piano courses [11], which could be lowered by supporting motivation and faster learning progress [36].

In this paper, we present P.I.A.N.O. to effectively support learning to play piano with interactive projection. In order to avoid the need for sight-reading and to reduce a learner's cognitive load, we developed an enhanced note visualization based on *piano roll notation* (see Fig. 1a). The notation originates from paper rolls used in the late 19th century, on which position and length of holes indicated notes' pitch and duration [7]. Using this approach in a graphical visualization allows the direct inference of the duration of notes (represented as vertical bars). Furthermore,

by projecting notes onto the piano, a note's pitch information (a bar's horizontal position) can be mapped directly to the respective piano key. Different colors of notes support correct fingering. We extended the basic piano roll notation with support for legato and staccato articulations, as well as trill and grace note ornaments (see Fig. 1b) as a first step towards fully visualizing sheet music.

P.I.A.N.O. provides three different modes to support the learning process (see Fig. 1c). In the *listen* mode, learners can listen to a song and follow its note visualization to become familiar with its rhythm and melody. The *practice* mode waits for correct key presses before advancing the note visualization. Learners are further supported by live feedback in finding correct keys. To evaluate the learning progress, the *play* mode measures the accuracy and errors of a performance. A detailed feedback screen facilitates identification of parts requiring further practice (see Fig. 1d).

We conducted two user studies to evaluate learning progress, required cognitive load, and user experience of learners when learning a song with P.I.A.N.O. compared to sheet music notation and basic piano roll notation without projection. The results of our within-subjects study (n_1=56) show that P.I.A.N.O. reduces the initial hurdle of starting to play piano by facilitating faster learning with less cognitive load. Results of a week-long between-subjects study with the three systems (n_2=18) show that P.I.A.N.O. facilitates playing with less errors and offers a steeper learning curve for novices practicing a song for one week (please note that the metaphor "steep learning curve" refers to a quick progress in learning throughout the paper). A qualitative rating of the final performances from the second study by 6 piano experts shows that P.I.A.N.O. also increases perceived quality and overall impression of the played music.

Contributions

The main contributions of our work are (1) results showing a significant reduction of cognitive load when learning piano playing with a projection-augmented piano that supports direct mapping of music notation to piano keys. (2) An enhanced piano roll notation conveying rich note information that does not require sight-reading and demonstrates the feasibility of visualizing complex sheet music notation. (3) Three learning modes (*listen*, *practice*, and *play*) that provide song preview, live feedback, and performance evaluation optimized for roll notation that effectively support the learning process. (4) Validation of the effectiveness of the aspects above with quantitative and qualitative evaluation in two user studies, and rating of perceived quality by piano experts.

We first summarize related work before describing the P.I.A.N.O. system in detail, followed by a presentation of its comparative evaluation. We conclude with a discussion of advantages, limitations, and potential extensions of our approach.

RELATED WORK

Music Games

Rhythm games, like *Guitar Hero* [1], reconcile a novice's desire to make music with speed of progress. In such games, colored notes scroll down the screen to a line of markers, which must be mapped to colored buttons on instrument-shaped controllers, and pressed in time to score points. Piano games use basic piano roll notation in a similar fashion. For instance, in the Android app *Pianist HD* [2], vertical bars of different length scroll towards a visual touch keyboard. The arcade game *Keyboardmania* [34] indicates pitch information by showing dots instead of bars on a display above a small keyboard similar to a real piano. The biggest drawback of those games is the use of controllers instead of real instruments and simulated music making, which provides entertainment to players but does not support learning to play real instruments. However, such games can nevertheless contribute to the development of some musical skills, such as visual tracking and rhythmic performance [12, 25].

Synthesia [30] merges rhythm games with actual piano learning on real instruments. A digital piano can be connected via a MIDI interface to a PC, which displays a basic piano roll notation of notes flowing towards a virtual keyboard. Scoring is calculated by correct pitch and duration of each note and can be shared online with other players. The high activity of the scoreboard[1] and the large number of tutorials for specific songs on Youtube[2] show that Synthesia is a widely adopted method for self-educated piano playing. However, players still need to map visual note representations to the instrument. P.I.A.N.O. projects notes directly onto the piano instead. Furthermore, Synthesia's basic piano roll notation provides no information about articulations or note ornaments.

Augmented Pianos

Basic approaches for augmenting a piano keyboard are *Disney's Piano Sound Book* [6] and the *Laugh & Learn Baby Grand Piano* [10] for preschool children. Here, light-up keys ease the mapping of a simplified sheet music notation to the keys of the toy piano. Casio also offers real digital pianos with light-up keys [4].

The main drawbacks of such approaches are that they only indicate the very next keys to play and that they rely only on sheet music notation, which does not support direct mapping of notes to keys [32, 23]. This direct mapping is supported in P.I.A.N.O. by projecting an enhanced piano roll notation directly onto the piano and its keys.

Direct Mapping of Notes to Keys

Toshio Iwai [17] artistically combined projection with direct note mapping to piano keys. Similar to the piano roll notation, light dots could be drawn as notes on a projected surface with a trackball. The light dots then moved towards the piano and were played automatically. While the projection method is similar to our approach, users could only draw dots and not directly play the piano. Yang and Essl [38] augmented a digital piano with a projection setup similar to P.I.A.N.O. They implemented different visualization methods, including a basic piano roll notation, to discuss their influence on the design space of an augmented piano. However, their visualization provided neither any information about articulation or fingering, nor performance feedback. P.I.A.N.O. significantly extends the projection approach by supporting these features and the learning process with interactive learning modes.

In contrast to our goal of supporting piano playing without sight-reading knowledge, Takegawa et al. [31] propose a projection-augmented piano to support learning of sheet music notation. Notes of the sheet notation are visually connected to piano keys

[1] `http://www.synthesiagame.com/scoreboard.aspx`
[2] 1,440,000 search results for "Synthesia" as of September 2014

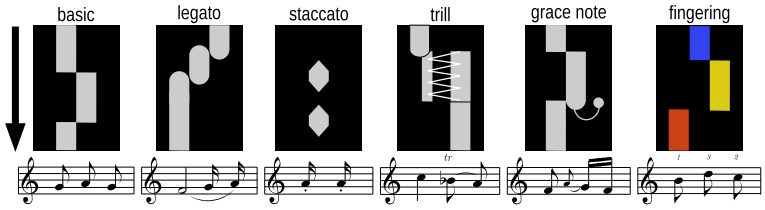

Figure 2. Different note and fingering representation in P.I.A.N.O. (*top*) compared to five-staff notation (*bottom*).

in order to support note to key mapping. The results of a between-subjects study (n=9) suggest that the method efficiently supports sheet music learning. However, the evidence is preliminary with only three learners per method and a short practice session of 30 minutes. Recently Raymaekers et al. [24] presented "The Augmented Piano" (TAP), with a setup similar to P.I.A.N.O.The system also provides a gamified alternative, wherein instead of note visualization the user shoots spaceships with the keys. However, no evaluation of their system was provided.

To the best of our knowledge, our work is the first to provide an extensive evaluation of an enhanced projected roll notation in a short-term study (n_1=56) and a one-week longitudinal study (n_2=18) to compare performance development between different learning systems. Furthermore, we do not aim to teach sight reading, but rather piano playing with a combination of our enhanced note visualization (with articulation marks and ornaments as a first step towards fully visualizing sheet music), correct fingering, and the three tightly coupled integrated learning modes optimized for roll notation.

Fingering Support

The *Concert Hands* [26] are finger sleeves worn during play, which signal correct fingering. Hands are also autonomously guided to the right position by a wrist pilot on a rail mounted in front of the piano. Huang et al. [15] propose a similar glove to passively learn correct fingering by indicated vibrations.

In Synthesia [30], correct fingering is supported by a color for each hand, and numbers (1–5) displayed with notes indicating which finger to use. However, numbers have to be assigned to notes manually beforehand, which requires learners to know which fingering approach is best suited for a song.

Xiao et al. [37] propose a remote piano tutoring system based on projected hands and autonomously moving keys of a player piano. Both student and teacher place a camera on top of the keyboard to capture hand positions and another one in front of the pianist to capture body language. Captured video is projected on to each others' piano. This way, piano teachers can remotely guide accurate fingering and body language.

P.I.A.N.O. uses different note colors to indicate correct fingering, similar to the *Glow Piano Lessons* app for iOS devices [16]. However, in contrast to the iOS app, only notes for important finger switches are colored, in order to draw attention to them without overstraining learners. While this approach requires a short learning phase of color-to-finger mapping, it facilitates fast fingering decisions during play.

THE P.I.A.N.O. SYSTEM

Rather than relying on sheet music, P.I.A.N.O. projects music notation directly onto the instrument. Notes are represented in piano roll notation, i.e., upcoming notes move towards the keyboard, which allows direct mapping of notes to the respective piano keys without any sight-reading knowledge. The current keys to play are illuminated to further ease this mapping. In contrast to basic roll notations common in music video games, our note visualization supports not only pitch and duration of notes, but also additional articulation marks, and recommended fingering, in order to provide similar information as conventional sheet music notation. The learning process itself is supported with different interactive modes following social learning theory [3].

Note Visualization

A major goal in the design of our note representation was to ease the learning process by shifting cognitive capacity from notation mapping to music playing. Traditional sheet music uses a static, symbolic notation, which is hard to match to the analogous actions required in playing the piano [32, 23]. McLachlan et al. [23] also found that spatial, graphical notations lead to improved performance in novices. Therefore, we represent notes with pictorial analogies to reduce extraneous cognitive load for the learner. We refined our note representation in an iterative process with many discussions within our research team (including two piano players and two band members with an average experience of 15 years in playing different instruments), as well as multiple practice sessions and semi-structured interviews with four participants [33].

In our system, upcoming notes are projected on the extended surface behind the keyboard to provide a preview (see Fig. 1a). On the keyboard itself, the current keys are highlighted to signal to the learner what to play. In our pre-study, many participants looked predominantly at the piano keys. Although this visual scope increased upwards with longer practice, we found a preview of upcoming notes on the keys to be very helpful as it indicates where hands have to move next. Therefore, upcoming notes are highlighted with thin extension lines, as shown in Figure 1 a) and b).

Figure 2 shows the main elements of our note visualization. The horizontal and vertical position of notes indicate what key to press (pitch) and when; their vertical length indicates the note's duration, i.e., how long the key has to be pressed. Shape and color of notes provide additional information. Our current visualization supports the articulations *legato* and *staccato*. Legato notes overlap and are extended by a half circle to indicate smooth playing without pause. In contrast, we encoded staccato with triangular ends to evoke "pointier" notes, which should be played

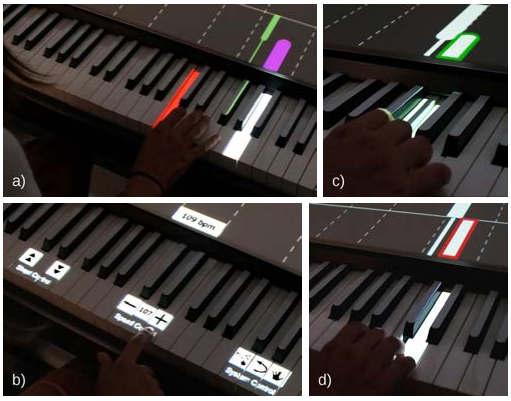

Figure 3. (a) In practice mode, an incorrectly played note is highlighted in red, the correct key in white. (b) The pedal-activated menu allows control of mode-specific settings, e.g., adapting the speed or scrolling through the visualization. (c) Live feedback in play mode highlights correctly played notes with green and (d) errors with red borders.

short and disjointedly. Common ornaments in sheet music are *trill* and *grace notes*. Trills are visualized with interconnecting lines between two notes to indicate fast repeated switching between them. A grace note is a short note played directly before the next one; it is indicated by a connected dot.

Fingering is indicated through colored notes with one color for each finger of the right hand. We chose the colors for their contrast between each other in order to prevent mistakes due to chromatic similarity. Our pre-study showed that fingering information for every note overwhelms learners. Therefore, similarly to sheet music, P.I.A.N.O. highlights only important finger switches. Notes without fingering information are gray and can be played with any suitable finger.

Learning Modes

Social learning theory [3] distinguishes four steps of learning. The first is *attention*; learners observe a process. In the *retention* step, learners try to remember what they observed. With *reproduction*, learners perform the observed behavior themselves. Further practice leads to improvement and skill advancement. *Motivation* consists of reinforcement and punishment to ensure that learners continue practice; this is essential for observational learning to be successful. P.I.A.N.O. supports social learning theory by offering three learning modes: *listen*, *practice*, and *play* (accessible from the main menu, see Fig. 1c). System and navigation controls are available by pushing down any of the piano's foot pedals. Mode-specific controls, e.g., adjusting speed, are then projected onto the keys and can be activated via keypress (see Fig. 3b).

Listen mode (attention)

In the *listen* mode, learners can listen to the song and follow its note visualization (*attention*). Thus, learners become acquainted with the characteristics of different song parts. This auditory learning aspect was revealed as highly important in conversations with experts (i.e., a professional piano teacher and a professor of psychology). Following social learning theory [3], the synchronized playback of auditory and visual information should increase learners' attention.

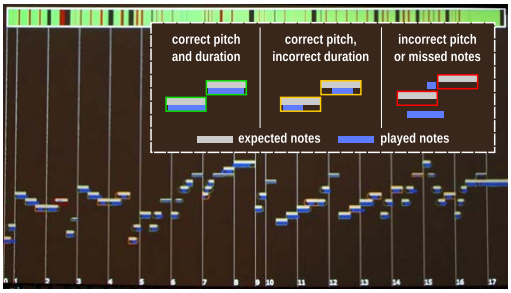

Figure 4. The play mode shows a detailed feedback screen. It shows notes played by the user (*blue bars*), notes played with correct pitch and duration (*green border*), correct pitch but incorrect duration (*yellow border*), and incorrect or missed notes (*red border*). The progress bar at the top provides a respective overview, with missed song parts in *black*.

Practice mode (retention)

The *practice* mode provides learners with an environment to practice accurate playing in order to support retention. The system encourages correct playing without haste by waiting for correct key presses and duration before continuing in the song. When a wrong note is played, the system brightly highlights the correct key and marks the wrongly-pressed key in red (see Fig. 3a). The pedal-operated context menu allows leaners to jump back and forth in the current song, change the speed or song part, or return to the main menu (see Fig. 3b).

Play mode (reproduction and motivation)

After practice, learners can assess their performance in the *play* mode by playing the song in one piece, hereby reproducing the previously observed and memorized behavior (*reproduction*). The speed of the song is adjustable to allow evaluation at various speeds. While playing, the learner receives with live feedback as part of the note visualization. A green border around the currently played note indicates a correctly played note; a red border indicates an incorrectly played note (see Fig. 3c & d). A progress bar at the top provides a summarized overview of correct (green), incorrect (red), and missed song parts (black). The progress bar is also shown in the final feedback screen (see Fig. 4), which provides detailed information of played notes (blue), and expected notes (grey). A colored border indicates correctly played notes (green), incorrectly played or missed notes (red), and notes played with correct pitch but incorrect duration (yellow). This feedback allows learners to identify which parts were played well and which require further practice, thus, it provides *motivation*.

Technical Aspects

We use a standard digital piano (Thomann DP-25), which supports MIDI communication. It is connected to a desktop computer (Intel Core i5-3470, 4 GB RAM, Intel HD Graphics 2500) via a MIDI/USB interface (Roland UM-One). The projection on the piano is created by a short-throw projector (BenQ MW516) displaying the computer's video output. It is mounted above the piano using a tripod. A smooth projection

Figure 5. Setup of the compared systems in the user studies: (a) P.I.A.N.O. (interactive projection), (b) Synthesia (piano roll notation), and (c) Finale (sheet music notation).

surface is provided by a wooden plate attached to the piano, covering its top and extending towards the back (see Fig. 5a).

The MIDI protocol identifies key presses and their duration: each key press and release generates an event message , including a corresponding timestamp. This allows us to compare the system's representation of a song's expected notes with notes played on the piano in order to provide live feedback. The internal representation of a song, used for evaluation as well as visualization, is generated by parsing song files in MusicXML format [22]. Thus, any song available in this format can be played with P.I.A.N.O.

In listen mode, the computer generates MIDI messages to control piano playback. In practice mode, played notes are evaluated to generate the colored hints indicating correctly or incorrectly pressed keys (see Fig. 3c & d). In play mode, live feedback is generated based on real-time information of pressed keys. In addition, all key presses and key releases are logged throughout a learner's performance. This log is then compared to the ideal performance generated by the system to yield the result screen displayed at the end (see Fig. 4).

EVALUATION

In order to examine P.I.A.N.O.'s impact on the learning performance of novices, we conducted two experimental studies. We evaluated the learning performance with the interactive projection of P.I.A.N.O. in comparison to a more game-orientated *roll notation* and the traditional way of learning the piano with *sheet music*. As specific comparison systems, we selected commercially available software: *Synthesia* [30], a popular educational piano game, and *Finale* [21], a music software with traditional sheet music notation. Using software for all three systems enabled us to evaluate performance with a consistent quantitative measurement approach.

The first study (n_1=56) employed a within-subjects design to compare initial performance, experienced cognitive load, and perceived user experience of the three systems. A between-subjects study followed (n_2=18) to assess learners' performance over one week. We were further interested in the systems' impact on perceived quality and overall impression of performances. Thus, final recordings of the second study were blindly and independently rated by 6 piano experts.

System Setup

As the three systems were identical for both studies, we first outline their setup and the employed quantitative measurement approach, before discussing the specifics of the studies. Figure 5 shows the experimental setup of each system. The P.I.A.N.O. system was set up as described above (see Fig. 5a). For Synthesia and Finale, we used a 24"-display which was placed in front of the learners (see Fig. 5b & c). Sound settings were consistent for all systems and sound output was directly generated by the piano's built-in speakers.

Synthesia [30] was chosen due to its popularity with self-tutoring piano learners. Synthesia visualizes notes in a basic piano roll notation. It offers a *melody practice* feature, similar to our practice mode, as well as a *song recital* feature that demands a defined speed similar to our play mode. Fingering information is limited in both its availability and visualization, and there is no visualization of articulation techniques. When connected to a piano via MIDI, Synthesia can assess a learner's performance from piano key presses.

Finale [21] is a software for composing and displaying sheet music. To achieve comparable conditions to our play mode between all systems, it was essential that the dynamic indication of the current position within a song was supported. Finale displays a dynamic marker that runs through the depicted sheet music at a defined speed. In addition, learners were offered a metronome to support keeping a song's rhythm. In contrast to P.I.A.N.O. and Synthesia, sheet music requires learners to perform sight-reading. Thus, learners were given additional help in form of a notation guide and an image mapping each note in staff notation to the keyboard. The piano's middle C note was marked with a colored sticker.

In order to obtain consistent quantitative performance results of all systems, we implemented a measurement tool that recorded the piano's MIDI output, i.e., actually pressed keys, in synchronization with the evaluated system. Similar to Drake and Palmer [8], recorded session logs were analyzed against expected correct notes: We differentiated between *missed notes*, *incorrectly pressed notes* (pitch errors) and *correctly pressed notes* (correct pitch). Incorrectly pressed notes were counted independently from the number of expected notes, i.e., pressing a wrong key twice during the duration of one expected note resulted in two incorrectly pressed notes. This allowed for a more precise distinction of incorrect performances. Correctly pressed notes were further sub-classified regarding their duration, i.e., *correct* and *incorrect duration notes*. Duration was evaluated based on absolute length comparison and relative timestamp matching.

User Study 1: Initial Performance and User Experience

In a within-subjects study with the three systems, we evaluated the performance of novices with no previous piano experience in learning to play the right-hand part of a song in 15 minutes. We expected that *learning with P.I.A.N.O. would result in higher learning performance* (more correctly pressed notes, less

incorrectly pressed and missed notes) (H_1) than Synthesia and Finale, due to the direct mapping of notes onto keys. For the same reason, we expected that *P.I.A.N.O. would create less intrinsic, less extraneous, and higher germane cognitive load* (H_2). *Intrinsic cognitive load* (ICL) results from the learning material itself, its element interactivity, and prior knowledge of the learner [29]. *Extraneous cognitive load* (ECL) occurs due to the design of the learning environment, whereas *germane cognitive load* (GCL) arises from the learner's intrinsic motivation and concentration on understanding the learning material. Hence, a system should strive for low ICL and ECL, but high GCL [29]. Finally, we expected that *P.I.A.N.O.'s user experience would be rated higher* (H_3) than those of the other systems.

Method and Material

With the help of two experienced pianists and a professional piano teacher, we selected three songs of similar difficulty and length (16 bars, 26 seconds) that have a dominant melody played with the right hand. The songs were chosen from a grade 1 syllabus of exam pieces from the Associated Board of the Royal Schools of Music [20]: "*Das Ballett*" by D. G. Türk, "*Minuet in G*" by W. A. Mozart, and "*Moderato*" by A. F. Gedike. To verify that the songs were of similar difficulty, we performed a treatment check based on learning performance with sheet music notation. We could not find a significant difference regarding correctly pressed notes. (song 1 vs. 2: *MD*=1.91, *SE*=4.40, n.s.; song 1 vs. 3: *MD*=3.62, *SE*=4.34, n.s.; song 2 vs. 3: *MD*=5.53, *SE*: 4.34, n.s.).

At the beginning of the session, participants completed a questionnaire regarding their demographic information, and musical background (i.e., experience with other instruments, and sight-reading skills). We further used a subtest of "the kit of factor-referenced cognitive tests" [9] to assess learners' spatial ability. Each participant interacted with all three systems and learned a different song with each. The order of systems and songs was counter-balanced (Latin square). For each system, participants were first given a written introduction to the specific system, and encouraged to try out the described features. Participants then listened to the assigned song three times (P.I.A.N.O.'s *listen* mode, Synthesia's *watch and listen only*, and Finale's *playback* mode). This was the only time when listening to the song was allowed. Playing performance was measured at four times: after listening to the song *without practice* (T_1), *after practicing the first half of the song (5 min.)* (T_2), *after practicing the second half of the song (5 min.)* (T_3), and a final measurement *after practicing the whole song (5 min.)* (T_4). However, each measurement corresponds to the learning performance for the whole song as measured in the system-specific play-through mode (P.I.A.N.O.'s *play* mode, Synthesia's *song recital*, and Finale's *dynamic-marker* mode).

After each system session, participants completed an 8-item survey [19], designed to differentiate between the three parts of cognitive load. Each item posed a statement to be rated on a 7-point Likert scale. Perceived user experience of each system was assessed with the AttrakDiff survey [14], which consists of 28 bipolar verbal anchors rated on a 7-point scale. AttrakDiff results in four scales: *perceived pragmatic quality (PQ)* measuring the support for achieving a goal, *hedonic quality – stimulation (HQ-S)* measuring perceived novelty and potential to grab users' attention,

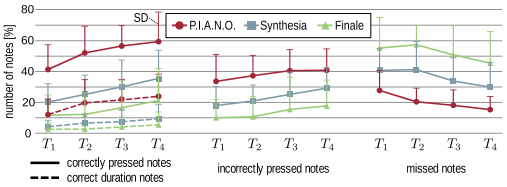

Figure 6. Learning performance for P.I.A.N.O., Synthesia, and Finale in first study for measurements T_1–T_4.

hedonic quality – identification (HQ-I) measuring potential of identification with the system, and *perceived attractiveness (ATT)*.

Sessions lasted about two hours on average, with 20-30 min. per system. Due to the long duration, participants were offered snacks and drinks, and afterwards allowed to choose from a variety of larger sweet collections as compensation.

Participants

A total of 56 right-handed learners participated in the study, all of them novices in playing piano. The average age of participants was 23 years (*SD*=3.92); they were fairly balanced in gender (32 female, 24 male), and most were students. Some participants had a basic level of ability to read sheet music; on average, participants achieved 6.98 out of 15 points in the test of sight-reading skills (*SD*=4.59).

Learning Performance

Figure 6 shows the results of the quantitative analysis of the learners' short-term learning performances. For all measurements (T_1–T_4), P.I.A.N.O. learners achieved a higher percentage of correct notes and also performed better regarding duration accuracy. An analysis of variance with repeated measures (rANOVA) revealed a significant difference in the number of *correctly pressed notes* ($F(2,88)=121.64$, $p<.001$, $\eta^2=0.73$) and *correct duration notes* ($F(1.37,61.85)=85.31$, $p<.001$, $\eta^2=0.66$) for all three systems. An analysis of contrasts showed that P.I.A.N.O. learners hit significantly more correct notes than Synthesia ($F(1,44)=120.42$, $p<.001$, $\eta^2=0.73$) and Finale learners ($F(1,44)=173.31$, $p<.001$, $\eta^2=0.80$), and also played more *correct duration notes* compared to Synthesia ($F(1,45)=78.04$, $p<.001$, $\eta^2=0.63$) and Finale ($F(1,45)=108.83$, $p<.001$, $\eta^2=0.71$).

With respect to *incorrectly pressed notes*, Figure 6 shows that P.I.A.N.O. learners also pressed more incorrect notes on average than learners of the other systems. An rANOVA revealed significant differences between all systems ($F(1.75,76.92)=60.11$, $p<.001$, $\eta^2=0.58$). The contrasts show that learning with P.I.A.N.O. results in more incorrectly pressed notes than learning with Synthesia ($F(1,44)=60.62$, $p<.001$, $\eta^2=0.58$) or Finale ($F(1,44)=85.70$, $p<.001$, $\eta^2=0.66$). However, on average, P.I.A.N.O. learners also tried to play more notes, which is reflected in the lower number of *missed notes*. Missed notes significantly differ between all systems (rANOVA: $F(1.73,77.79)=89.83$, $p<.001$, $\eta^2=0.67$) and P.I.A.N.O. learners missed significantly less notes than Synthesia learners ($F(1,45)=72.30$, $p<.001$, $\eta^2=0.62$) and Finale learners ($F(1,45)=165.04$, $p<.001$, $\eta^2=0.79$). In summary, with the exception of incorrectly pressed notes, the

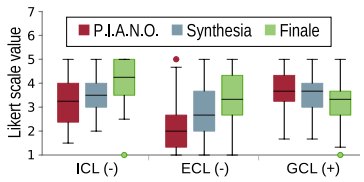

Figure 7. Intrinsic, extrinsic, and germane (positive) cognitive load of P.I.A.N.O., Synthesia, and Finale in study 1.

combined results confirm our hypothesis that P.I.A.N.O. improves initial learning performance (H_1).

Cognitive load

Figure 7 shows the cognitive load results. Internal consistency (Cronbach's α) of items was confirmed for intrinsic cognitive load (ICL) (P.I.A.N.O.: α=.81; Synthesia: α=.79; Finale: α=.70), extrinsic cognitive load (ECL) (P.I.A.N.O.: α=.81; Synthesia: α=.86; Finale: α=.87), and germane cognitive load (GCL) (P.I.A.N.O.: α=.62; Synthesia: α=.53; Finale: α=.58).

According to an rANOVA, significant differences existed for ICL ($F(2,108)=18.39$, $p<.001$, $\eta^2=0.25$), ECL ($F(2,108)=16.88$, $p<.001$, $\eta^2=0.24$), and GCL ($F(2,108)=10.02$, $p<.001$, $\eta^2=0.16$). P.I.A.N.O. induces significantly lower ICL than both Synthesia ($F(1,54)=5.02$, $p<.05$, $\eta^2=0.09$), and Finale ($F(1,54)=38.00$, $p<.001$, $\eta^2=0.41$), indicating that the complexity of the note notation is reduced by P.I.A.N.O.'s projected roll notation compared to Synthesia or Finale. ECL was also rated significantly lower with P.I.A.N.O. than with Synthesia ($F(1,54)=9.64$, $p<.01$, $\eta^2=0.15$) or Finale ($F(1,54)=28.71$, $p<.001$, $\eta^2=0.35$). This suggests that P.I.A.N.O. reduces split attention: In Synthesia, attention is split between the keyboard and the display. The metronome may further split attention in Finale. P.I.A.N.O. learners only needed to follow the projected notation which is seamlessly integrated with the keyboard.

Furthermore, P.I.A.N.O. induces significantly more positive GCL than Finale ($F(1,54)=18.59$, $p<.001$, $\eta^2=0.26$). We could not find a significant difference between P.I.A.N.O. and Synthesia ($F(1,54)=2.88$, n.s.); presumably their roll notations promote similar motivation. Thus, H_2 can be accepted, as P.I.A.N.O. induces lower intrinsic and extrinsic load than Synthesia and Finale, and higher germane load than Finale.

User experience

The per-item ratings of the three systems on the AttrakDiff scales are shown in Figure 8. P.I.A.N.O. was ranked highest in almost all cases, with Finale consistently ranked lowest. Significant differences exist for perceived pragmatic quality (PQ) ($F(2,108)=19.97$, $p<.001$, $\eta^2=0.25$), stimulation (HQ-S) ($F(2,108)=117.76$, $p<.001$, $\eta^2=0.69$), and attractiveness (ATT) ($F(2,108)=24.75$, $p<.001$, $\eta^2=0.31$). Differences were not significant for the identification scale (HQ-I). More specifically, for all scales P.I.A.N.O. was rated significantly higher than Synthesia (PQ: $F(1,54)=8.64$, $p<.01$, $\eta^2=0.15$; HQ-S: $F(1,54)=46.78$, $p<.001$, $\eta^2=0.46$; ATT: $F(1,54)=11.85$, $p<.001$, $\eta^2=0.18$), and Finale (PQ: $F(1,54)=28.26$, $p<.001$, $\eta^2=0.34$; HQ-S: $F(1,54)=225.33$, $p<.001$, $\eta^2=0.81$; ATT: $F(1,54)=39.50$, $p<.001$, $\eta^2=0.42$). As a result, H_3 can also be accepted.

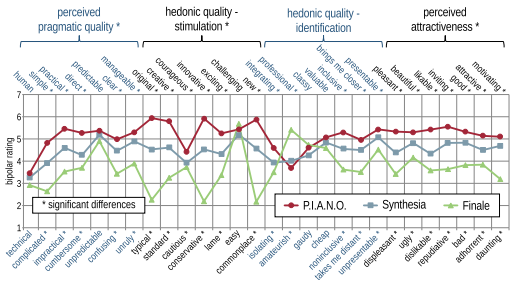

Figure 8. User experience results rated with AttrakDiff's 28 bipolar verbal anchors.

User Study 2: One-week Performance

In the first study, P.I.A.N.O. and its interactive projection outperformed classic piano roll notation (Synthesia) and sheet music notation (Finale) for most of the quantitative performance metrics, cognitive load, and user experience. These results suggest that projecting an enhanced note visualization onto the piano significantly eases piano learning for novices. In order to validate P.I.A.N.O.'s superiority regarding performance (correct, incorrect, and missed notes), we conducted a second study with the primary goal of evaluating learner performance over a longer period of time. We recruited a new group of participants (n_2=18) that practiced with one of the systems each day for one week. We opted for a between-subjects design to avoid performance interference from multiple systems. Our hypothesis was that *P.I.A.N.O. would retain a steeper learning curve over one week than Synthesia or Finale*, based on measured performance (H_4).

Method and Material

In order to avoid a ceiling effect, the song for Study 2 was chosen through a preliminary study with 3 participants in order to select a song of which the right-hand part was neither too easy nor too hard to learn. Based on the results, we chose a slightly simplified version of Schumann's *Träumerei* (16 bars, 116 notes), wherein the difficulty consisted mostly of legato articulation and several two- and three-note chords. Three groups, each consisting of six participants were monitored while they practiced with one system (P.I.A.N.O., Synthesia, or Finale) for a total of 5 consecutive days. Each daily session consisted of a brief questionnaire regarding their pre-session physical condition, followed by listening to the song one time, a 15-minute practice session, and a daily play-mode measurement (D_1–D_5).

Participants

A total of 18 learners participated in this study; mainly students, 26 years old on average (SD=4.45), and balanced in gender (9 female, 9 male), which were evenly distributed across the three groups. The preliminary sight-reading test revealed no differences in the average sight-reading skills among learners of P.I.A.N.O. (M=8.8, SD=5.42), Synthesia (M=9, SD=4.69), and Finale (M=8.3, SD=4.32).

Learning Performance

The participants' learning curves for the three systems are depicted in Figure 9. For all measurements, P.I.A.N.O. learners achieved a higher percentage of correct notes, with better duration

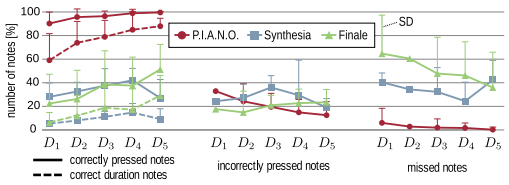

Figure 9. Development of learning performance in second study for 5 consecutive days (D_1–D_5).

accuracy (on average 88% *correct duration notes* on the last day) and lower percentage of *incorrectly played* and *missed notes* (12% and 0.3% on last day). An rANOVA revealed significant differences in the learning performance regarding *correctly pressed notes* ($F(2,15)=34.26$, $p<.001$, $\eta^2=0.82$), *correct duration notes* ($F(2,15)=51.43$, $p<.001$, $\eta^2=0.87$), and *missed notes* ($F(2,15)=12.41$, $p<.01$, $\eta^2=0.62$) at all five measurements. A significant interaction of system and measurement point was found, showing a steeper learning curve of P.I.A.N.O. learners regarding *correctly pressed notes* ($F(3.62,3)=5.15$, $p<.01$, $\eta^2=0.41$), *correct duration notes* ($F(4.81,3)=4.24$, $p<.01$, $\eta^2=0.36$), and *missed notes* ($F(3.50,3)=3.59$, $p<.05$, $\eta^2=0.32$) compared to those of Synthesia and Finale learners. For *incorrectly pressed notes*, an rANOVA showed no significant differences between systems ($F<1$, n.s.) or measurements ($F(2.00,13)=1.88$, n.s.). However, a difference in interaction was found ($F(4.00,3)=3.69$, $p<.05$, $\eta^2=0.33$); *incorrectly pressed notes* decreased over time for P.I.A.N.O. learners, were unstable for Synthesia learners, and increased for Finale learners. Together the results for *correctly/incorrectly pressed notes*, *correct duration notes*, and *missed notes* confirm our hypothesis that P.I.A.N.O. leads to an improvement in learning performance over time (H_4).

Expert Evaluation: Perceived Quality

6 piano experts were recruited (4 piano teachers, 2 professional players) with 28 years of experience on average, in order to gain qualitative performance ratings. We expected that *experts would rate recordings of P.I.A.N.O. learners higher than those of Synthesia and Finale learners in terms of perceived quality and overall impression* (H_5).

Method and Material

On the final day of study 2, we recorded one *practice* and one *play* session of each participant, resulting in 36 recordings (12 for each system). *Play* sessions were used for the expert evaluation because they provided consistent tempo, which is necessary for a realistic and comparable scenario. However, in some cases the play mode resulted in incomplete recordings when learners were unable to recover from playing mistakes. *Practice* sessions, therefore, allowed us to provide experts with more complete recordings for their ratings.

Each expert listened to all 36 recordings and filled out a survey for each. Recordings were anonymized regarding the system and playing mode. The three evaluated systems were presented to the experts in randomized order to avoid fatigue and assimilation effects. The rating criteria in the survey were a subset from a music performance adjudication set created by Wrigley et al. [35], and consisted of *pitch accuracy*, *duration accuracy*, *tempo*, *rhythm*, *continuity*, *dynamic*, *articulation*, *interpretation*, and *overall impression*. Each criterion had to be rated on a 7-point Likert scale (from *very poor* to *very good*). Furthermore, experts could comment on positive and negative aspects about the session in a text field. As a reward, each expert received a 10-Euro shopping coupon.

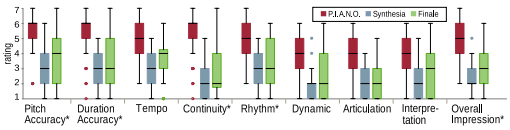

Figure 10. Expert ratings of perceived quality for 12 recordings (6 practice, 6 play) of each system.

Results

The ratings for associated play and practice sessions were combined by calculating their means. Calculated inter-rater reliability (Krippendorff's α) showed reliable ratings for 5 of the 9 scales: *pitch accuracy* (α=0.68), *duration accuracy* (α=0.66), *rhythm* (α=0.69), *continuity* (α=0.61), and *overall impression* (α=0.70). P.I.A.N.O. recordings were rated higher for all scales compared to recordings of Synthesia and Finale (see Fig. 10). For the reliable scales, ANOVAs revealed significant differences in pitch accuracy ($F(2,15)=24.30$, $p<.001$, $\eta^2=0.76$), duration accuracy ($F(2,15)=11.92$, $p<.01$, $\eta^2=0.61$), continuity ($F(2,15)=15.61$, $p<.001$, $\eta^2=0.68$), rhythm ($F(2,15)=14.47$, $p<.001$, $\eta^2=0.66$), and the overall impression ($F(2,15)=13.67$, $p<.001$, $\eta^2=0.65$). Thus, we can confirm our hypothesis H_5 for these scales.

DISCUSSION

Both user studies revealed significant differences in multiple variables between the three systems. P.I.A.N.O. learners were able to play more correct notes upon first play, and improved their performance more notably regarding duration accuracy, incorrectly pressed notes (pitch errors), and missed notes. The group of piano experts further confirmed that performances of P.I.A.N.O. learners achieved higher overall impression and perceived quality after one week of practice. Additionally, learning with P.I.A.N.O. resulted in less negative intrinsic and extrinsic cognitive load, more positive germane cognitive load, and provided a better user experience.

These results confirm the effectiveness of our projected piano roll notation, which allows direct mapping of notes onto the keys and does not require sight reading skills. Learners are able to concentrate on playing respective keys correctly (in terms of duration and basic articulation), rather than having to translate sheet music notation and find the correct keys. While trained pianists process mapping of sheet music notation automatically, it is a burdensome challenge for novices [28] and often leads to frustration [11, 36] – as also indicated by the lower average germane load measured for Finale. As one Finale learner commented: *"Without sight-reading skills, this task is very complex and absolutely frustrating."* P.I.A.N.O. eliminates this hurdle and allows song playing instantly. Most P.I.A.N.O. learners commented that time went by very quickly and that they had fun: *"Great when you want to learn playing piano in 5 minutes"*; *"very innovative and intuitive system which is fun to use."* In summary, learning piano with interactive projection appears to be easier and more fun than learning with sheet music notation or piano roll notation without direct mapping.

Limitations and Future Work

The study participants were drawn mainly from the current and former student population, and thus provide a fairly consistent age and educational background. Further studies are needed to investigate different populations, e.g., children or older adults. The second user study took place over the course of one week. While the measured learning curves proved highly promising for P.I.A.N.O., their further development over several weeks, months or even years is also of interest and might provide further valuable insights.

P.I.A.N.O. currently supports only right-hand song parts to target an early stage of learning. In order to provide support for more advanced learners and learning progression, we plan to extend our system to accommodate left-hand parts, as well as advanced features of conventional sheet music (e.g., crescendo or fermata). These visualization features could dynamically appear based on learners' performance. The system should also support fingering recognition, i.e., determining whether the learner pressed the key with the correct finger. We are currently integrating fingering recognition into our system, and plan to evaluate its influence on learning performance in a longitudinal study over several weeks.

Some learners were concerned about being too dependent on the system, and the lack of sheet music notation: *"Without the system I would not be able to play the song."* and *"You do not learn sight-reading."* While it was not our goal to support sight-reading skills, roll notation is unlikely to efficiently support the development of such skills. However, most learners in our studies were not interested in learning sheet music.

A common concern of consulted piano teachers was that the projected roll notation may lead to too much focus on correctness ("stiff playing"), thus hindering the development of richer musical expression. However, the results of our expert evaluation seem to refute those concerns. Although some qualitative scales (e.g., articulation and interpretation) did not provide significant differences, they do show higher tendencies. Some experts were initially concerned about rating novices who practiced for only one week, but were positively surprised (without knowing the used system) by recordings performed by P.I.A.N.O. players. This is also reflected in the significantly higher ratings on the overall impression scale. Although P.I.A.N.O. does not directly teach advanced musical expression, it could support learners in focusing on musical expression rather than on mapping notes to keys, by avoiding the burden of sight-reading. We plan to conduct extended longitudinal studies to examine whether these assumptions hold for practice spanning several weeks or months.

Developing musical expression is a prolonged process and playing correct notes is essential before proceeding to advanced skills. Our studies show that our approach supports quick success in terms of correctness at the first stage of piano playing. This success is essential for beginners when learning an instrument, as it motivates learners, could lower dropout rates and could even encourage people to start playing piano. While P.I.A.N.O. does not intend to replace individualized advice by expert piano teachers, it provides a significant improvement over current self-tutoring approaches. Piano teachers should further guide players with detailed advice, e.g., on fingering techniques and musical expression. This valuable advice remains irreplaceable by a system, although systems like MirrorFugue [37] try to support giving such advice remotely. In contrast, our aim was to support novices without sight-reading knowledge in faster learning and playing, focusing on developing basic playing skills and articulation. Our results show that this goal has been achieved, highlighting the validity of our chosen approach.

CONCLUSION

In this paper, we introduced P.I.A.N.O. as a novel approach to learning piano without sight-reading skills. Our enhanced piano roll notation is capable of depicting a note's pitch and duration, as well as a variety of articulation techniques (legato, staccato, grace notes, and trill notes). The projected notation allows a direct mapping of notes to respective keys, while three different modes further support the learning process. We conducted two user studies to compare the learning performance of P.I.A.N.O. to Synthesia and Finale, which use a basic piano roll notation without direct mapping and traditional sheet music, respectively. The results show that P.I.A.N.O. induces less cognitive load, better supports initial learning performance and faster progress over one week of practice, provides better user experience and leads to better perceived quality than Synthesia and Finale. Based on our results, we argue that the projected roll notation is a viable alternative to sheet music notation for beginners looking for a rewarding approach towards learning to play the piano.

ACKNOWLEDGMENTS

The authors would like to thank all study participants, the anonymous reviewers for their valuable feedback. This work was partially supported by the Transregional Collaborative Research Centre SFB/TRR 62 "Companion-Technology of Cognitive Technical Systems" funded by the German Research Foundation (DFG).

REFERENCES

1. Activision Publishing, Inc. Guitar Hero, accessed: Jun. 2014. `http://guitarhero.com/`.
2. Android appliaction. Pianist HD, accessed: Jun. 2014. `https://play.google.com/store/apps/details?id=com.rubycell.pianisthd`.
3. Bandura, A., and McClelland, D. C. Social learning theory. *General Learning Press* (1977).
4. Casio. Lighted keys keyboard LK-280, accessed: Jun. 2014. `http://www.casio.com/products/Digital_Pianos_&_Keyboards/Lighted_Keys/LK-280/`.
5. Creech, A., Hallam, S., McQueen, H., and Varvarigou, M. The power of music in the lives of older adults. *Research Studies in Music Education 35*, 1 (2013).
6. Disney. *Mickey Mouse Clubhouse Piano Sound Book: Mickey's Piano Party*. Publications International, 2009.
7. Dolan, B. *Inventing entertainment: the player piano and the origins of an American musical industry*. Rowman & Littlefield Publishers, Lanham, Md., 2009.
8. Drake, C., and Palmer, C. Skill acquisition in music performance: Relations between planning and temporal control. *Cognition 74*, 1 (2000), 1–32.
9. Ekstrom, R. B., French, J. W., and Harman, H. H. *Manual for kit of factor-referenced cognitive tests*. Educational Testing Service, Princeton N.J., 1976.

10. Fisher Price. Laugh & Learn Baby Grand Piano, accessed: Jun. 2014. http://www.fisher-price.com/en_AU/brands/babytoys/products/40014.
11. Giomi, E. C. "I do not want to study piano": early predictors of student dropout behavior. *Bulletin of the Council for Research in Music Education*, 161 (2004).
12. Gower, L., and McDowall, J. Interactive music video games and children's musical development. *British Journal of Music Education 29*, 1 (2012), 91–105.
13. Hallam, S. The power of music: Its impact on the intellectual, social and personal development of children and young people. *International Journal of Music Education 28*, 3 (2010), 269–289.
14. Hassenzahl, M., Burmester, D. M., and Koller, F. AttrakDiff: a questionnaire to measure perceived hedonic and pragmatic quality. In *Mensch & Computer*, Springer (2003).
15. Huang, K., Starner, T., Do, E., Weiberg, G., Kohlsdorf, D., Ahlrichs, C., and Leibrandt, R. Mobile music touch: mobile tactile stimulation for passive learning. In *Proc. CHI '10*, ACM (2010).
16. iOS application. Glow Piano Lessons, accessed: Jun. 2014. https://itunes.apple.com/us/app/glow-piano-lessons/id441970188.
17. Iwai, T. Piano–as image media. *Leonardo 34*, 3 (2001).
18. Jutras, P. J. The benefits of adult piano study as self-reported by selected adult piano students. *Journal of Research in Music Education 54*, 2 (July 2006), 97–110.
19. Klepsch, M., and Seufert, T. Subjective Differentiated Measurement of Cognitive Load. In *Proc. 5th Intl. Cognitive Load Theory Conf.* (2012).
20. London-Based Examinations Board. Associated Board of the Royal Schools of Music (ABRSM), accessed: Jun. 2014. http://de.abrsm.org/en/home.
21. MakeMusic, Inc. Finale - music notation software, accessed: Jun. 2014. http://www.finalemusic.com/.
22. MakeMusic, Inc. MusicXML, accessed: Jun. 2014. http://www.musicxml.com/.
23. McLachlan, N. M., Greco, L. J., Toner, E. C., and Wilson, S. J. Using spatial manipulation to examine interactions between visual and auditory encoding of pitch and time. *Frontiers in Psychology 1* (2010).
24. Raymaekers, L., Vermeulen, J., Luyten, K., and Coninx, K. Game of tones: Learning to play songs on a piano using projected instructions and games. In *Proc. CHI EA '14*, ACM (2014), 411–414.
25. Richardson, P., and Kim, Y. Beyond fun and games: A framework for quantifying music skill developments from video game play. *Journal of New Music Research 40*, 4 (2011), 277–291.
26. Rubato productions. Concert Hands, accessed: Jun. 2014. http://www.concerthands.com/products.html.
27. Sloboda, J. A., Clarke, E. F., Parncutt, R., and Raekallio, M. Determinants of finger choice in piano sight-reading. *Journal of Experimental Psychology: Human Perception and Performance 24*, 1 (1998), 185–203.
28. Stewart, L., Walsh, V., and Frith, U. Reading music modifies spatial mapping in pianists. *Perception & psychophysics 66*, 2 (2004), 183–195.
29. Sweller, J., Van Merrienboer, J. J., and Paas, F. G. Cognitive architecture and instructional design. *Educational Psychology Review 10*, 3 (1998), 251–296.
30. Synthesia LLC. Synthesia - piano game, 2013. http://www.synthesiagame.com/.
31. Takegawa, T., Terada, T., and Tsukamoto, M. A piano learning support system considering rhythm. In *Proc. ICMC '12*, Michigan Publishing (2012).
32. Tan, S.-L., Wakefield, E. M., and Jeffries, P. W. Musically untrained college students' interpretations of musical notation: sound, silence, loudness, duration, and temporal order. *Psychology of Music 37*, 1 (2009), 5–24.
33. Weing, M., Röhlig, A., Rogers, K., Gugenheimer, J., Schaub, F., Könings, B., Rukzio, E., and Weber, M. P.I.A.N.O.: Enhancing Instrument Learning via Interactive Projected Augmentation (demo). In *UbiCom '13 Adjunct Proceedings* (2013).
34. Wikipedia. Keyboardmania. version ID: 542589199, 2013. http://wikipedia.org/wiki/Keyboardmania.
35. Wrigley, W. J., and Emmerson, S. B. Ecological development and validation of a music performance rating scale for five instrument families. *Psychology of Music 41*, 1 (2013), 97–118.
36. Wristen, B. Demographics and motivation of adult group piano students. *Music Education Research 8*, 3 (2006), 387–406.
37. Xiao, X., Pereira, A., and Ishii, H. MirrorFugue III: conjuring the recorded pianist. In *Proc. NIME'13* (2013).
38. Yang, Q., and Essl, G. Visual associations in augmented keyboard performance. In *Proc. NIME '13* (2013).

Klepsch, M., Könings, B., Weber, M. & Seufert, T. (25. August 2014). Fostering Piano Learning by dynamic mapping of notes. European Association for Research on Learning and Instruction SIG 2 (Hrsg.), *EARLI SIG 2 Meeting "Building bridges: Improving our understanding of learning from text and graphics by making the connection"*, Rotterdam (NE).

Fostering Piano Learning by dynamic mapping of notes

Melina Klepsch
Ulm University, Institute of Psychology & Education
Albert-Einstein-Allee 47, 89081 Ulm, Germany
melina.klepsch@uni-ulm.de

Bastian Könings
Ulm University, Institute of Media Informatics
James-Frank-Ring, 89081 Ulm, Germany
bastian.koenings@uni-ulm.de

Michael Weber
Ulm University, Institute of Media Informatics
James-Frank-Ring, 89081 Ulm, Germany
micheal.weber@uni-ulm.de

Tina Seufert
Ulm University, Institute of Psychology & Education
Albert-Einstein-Allee 47, 89081 Ulm, Germany
tina.seufert@uni-ulm.de

Abstract. Learning to play piano is demanding because one has to map music notations from a sheet onto piano keys. Moreover learners have to look back and forth between the sheet and the piano. Overall learning performance will be low and cognitive load should be increased. To assist piano learners with these demands P.I.A.N.O., a piano learning system with interactive projection of music notation was developed. P.I.A.N.O directly projects visual hints upon the instrument and allows mapping of notes to piano keys without any sight-reading skills. The hint indicates which key has to be pressed and for how long. In an experimental study we analysed whether learners show better performance and less cognitive load after training with P.I.A.N.O compared to Finale, a software system with a dynamic marker running through the depicted notes in sheet notation. The P.I.A.N.O system in fact turned out to be highly effective.

Keywords: multimedia; split attention; information mapping; learning piano

Introduction and Theoretical Background

Learning to play piano is a cognitively demanding task. The first requirement is that learners have to understand musical notations. Secondly one has to map these signs onto the piano keys which is not supported by the structure of the conventional music sheet (McLachlan, Greco, Toner, & Wilson, 2010). While pitch is depicted by a vertical difference piano keys represent lower or higher pitches horizontally. Also duration and articulation of notes while playing have to be mapped from complex note symbols on the sheet to timing and pressure on keys. Overall this process should impose high cognitive load, especially high extraneous cognitive load due to visual search processes that hinder learners to concentrate on playing itself.

To assist learners with these demands, P.I.A.N.O. – an interactive projected roll notation system - was developed (Weing et al, 2013). The first aim was to avoid the need for sight-reading. Hence, notes are dynamically projected as visual signs on a roll notation system (Fig. 1). This system resembles those paper rolls used in 19th century, on which holes with their specific position and length indicated pitch and duration of notes to be played (Dolan, 2009). Notes are directly projected onto the piano so that learners do not have to split their attention from the sheet to the keys and the mapping process from notes to keys is effectively assisted. After an introduction phase where one can listen to the music and follow the note visualization learners can practice and receive feedback on accuracy and errors. A high score is given as a motivating issue.

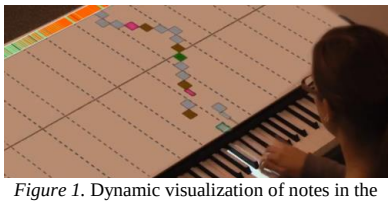

Figure 1. Dynamic visualization of notes in the P.I.A.N.O system

We conducted a study to evaluate the learning progress and cognitive load when learning with P.I.A.N.O. compared to the system *Finale* (MakeMusic, Inc., 2013), a pure sheet notation program.

Finale is a software to display sheet music with a dynamic marker that runs through the depicted notes. The information can be seen on a screen separate to the piano keyboard.

We expect, learning with P.I.A.N.O. would result in higher learning performance than learning with Finale (H1) due to the direct mapping of notes upon keys and the fact that no sight-reading skills are required.

Based on cognitive load theory (Chandler & Sweller, 1991) we also argue that cognitive load should be optimized by using P.I.A.N.O. (H2): We assume the projected roll notation in P.I.A.N.O. should reduce intrinsic cognitive load (ICL) by displaying less interrelated and complex elements like notes. With P.I.A.N.O learners only have to deal with one integrated learning material and there is no need to split attention between sheet music on a separated screen (Finale) and piano keys, i.e. the extraneous cognitive load (ECL) should be reduced. Germane cognitive load (GCL) is assumed to be higher when learning with P.I.A.N.O. than learning with Finale as the gamification through the feedback screen should result in higher engagement.

Method

In the experimental study 56 right-handed learners participated (average age: 23.05 years, SD = 3.92); 32 of them were female). All of them were novices in playing piano and their sight-reading skills were on a middle level with M = 6.98 out of 15 points (SD = 4.59).

Three songs of similar difficulty and length (16 bars, 26 seconds) which have a dominant melody played with the right hand were selected. At the beginning of a session, participants completed a questionnaire regarding demographics, experience with other instruments and sight-reading skills.

We conducted a within-subject design with the training system as independent factor: Each participant interacted with the two systems, P.I.A.N.O and Finale and in each case learned to play a different song. The order of systems and songs was counter-balanced.

After an introduction for each system learners listened to the assigned song three times. Playing performance as dependent variable was measured at four times, each 5 minutes of practice. After each system session (20-30 min.), participants completed an 8-item survey, designed to differentiate between the three parts of cognitive load as the second dependent variable (Klepsch & Seufert, 2012) with a 7-point Likert scale. Cronbachs Alpha for the differential scales were sufficient (between .62 and .87), except for germane load measured after using the Finale system (α=.58).

Results

Learning Performance

An analysis of variance with repeated measures (rANOVA) revealed a significant difference between the two systems in the number of correctly pressed notes ($F(1,48)=170.18$, $p<.001$, $\eta^2=0.78$) over all four measurements (see Figure 2).

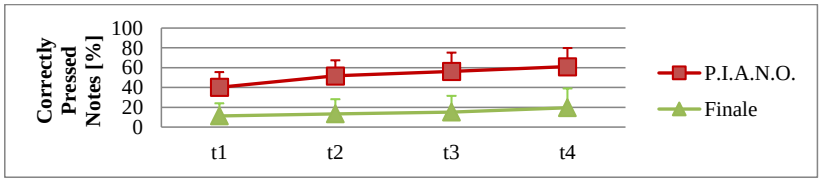

Figure 2. Learning performance over time for P.I.A.N.O. and Finale.

Moreover, there is a significant main effect for the repeated measure indicating a learning progress overall ($F(3,144)=38.5$, $p<.001$, $\eta^2=0.45$) which is especially considerable for P.I.A.N.O (fast success and stronger improvement) as the interaction effect demonstrates ($F(2.495,119.756)=5.98$, $p=.002$, $\eta^2=0.11$).

Cognitive load

According to a rANOVA, there are main effects for ICL ($F(1,54)=38.0$, $p<.001$, $\eta^2=0.41$), ECL ($F(1,54)=28.71$, $p<.001$, $\eta^2=0.35$), and GCL ($F(1,54)=18.59$, $p<.001$, $\eta^2=0.26$).

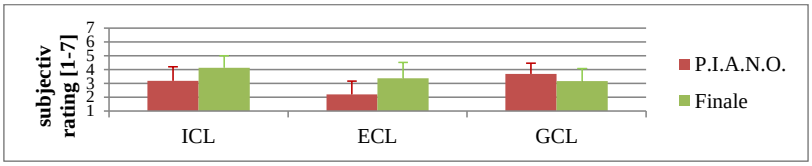

Figure 3. Subjective differentiated cognitive load rating for P.I.A.N.O and Finale.

Summary and Discussion

The study revealed several significant differences between the two systems. Learners were faster successful with P.I.A.N.O. and overall were able to play more correct notes. Additionally, learning with P.I.A.N.O. resulted in less ICL and ECL and enhanced GCL. These results confirm the effectiveness of the projected roll notation, which allows direct mapping of notes on to keys. Learners are able to concentrate on pressing keys correctly, rather than having to translate sheet music notation and find correct keys. Hence their cognitive resources are effectively used for learning instead of searching processes. While the measured learning curves proved to be highly promising for P.I.A.N.O., their further development over longer periods as well as motivation and users experience might provide further valuable insights.

References

Chandler, P., & Sweller, J. (1991). Cognitive Load Theory and the Format of Instruction. *Cognition & Instruction*, *8*(4), 293–332.

Dolan, B. (2009). *Inventing entertainment: The player piano and the origins of an American musical industry*. Lanham, Md: Rowman & Littlefield Publishers.

Klepsch, M., & Seufert, T. (2012). Subjective Differentiated Measurement of Cognitive Load. In *Proc. 5th Intl. Cognitive Load Theory Conf* .

MakeMusic, Inc. (2013). Finale - The World Standard in Music Notation Software. Retrieved from http://www.finalemusic.com/

McLachlan, N. M., Greco, L. J., Toner, E. C., & Wilson, S. J. (2010). Using Spatial Manipulation to Examine Interactions between Visual and Auditory Encoding of Pitch and Time. *Frontiers in Psychology*, *1*. doi:10.3389/fpsyg.2010.00233

Weing, M., Röhlig, A, Rogers, K., Gugenheimer, J., Schaub, F., Koenings, B., Rukzio, E. & Weber, M. (2013). P.I.A.N.O.: Enhancing Instrument Learning via Interactive Projected Augmentation. *Proceedings of the ACM conference on Pervasive and ubiquitous computing adjunct publication, UBiComp 13*, 75-78.

Teil III.

Diskussion und Ausblick

Man bleibt jung, solange man noch lernen, neue Gewohnheiten annehmen und Widerspruch ertragen kann.

(Marie von Ebner-Eschenbach)

5. Diskussion

Die vorliegende Arbeit befasst sich mit der theoriegeleiteten Entwicklung eines Fragebogens zur differenzierten Messung der kognitiven Belastung beim Lernen. Ziel war, das der entwickelte Fragebogen differenziert alle Arten kognitiver Belastung, die für die Beantwortung von Instruktionsdesignfragestellungen wichtig sind, erfasst. Basierend auf dem drei-Faktoren-Modell der CLT gehören dazu drei Arten der kognitiven Belastung: (1) Der ICL der von der Komplexität des Lerninhaltes abhängt. (2) der ECL welcher von der Darstellung des Lernmaterials beeinflusst wird. Und (3) der GCL der von der Aktivierung der kognitiven Prozesse der Lernenden abhängt. Diese Sicht widerspricht insgesamt nicht dem zwei-Faktoren-Modell der CLT, in welchem der GCL lediglich einen anderen Stellenwert (und Namen) erhält, für Instruktionsdesignfragestellungen jedoch weiterhin von Relevanz ist.

Neben dem Entwicklungsziel (Klepsch et al., 2017; Klepsch & Seufert, 09.-11.04.2012; Pichler & Seufert, 2011), war auch die Validierung und Testung des Fragebogens in weiteren Studien ein Ziel der Dissertation, welches durch die Durchführung von mehreren Studien erreicht werden konnte. Die meisten davon waren darauf ausgelegt explizit genau eine Art von kognitiver Belastung zu variieren um die Ergebnisse auch genau auf diese Variation beziehen zu können (Klepsch & Seufert, 2020; Seufert et al., 2017; Klepsch, Kempter & Seufert, 2013a, 2013b). Im praktischen Einsatz, z.B für eigene Lehrmaterialien oder eine Lernumgebung ist dies natürlich nicht gegeben. Die P.I.A.N.O.-Studie zeigt genau dieses Einsatzszenario (Rogers et al., 2014; Klepsch et al., 25.-27.08.2014). Eine Lernumgebung wurde entwickelt und soll mit der ursprünglichen Lernumgebung verglichen werden. Hierbei wurden jedoch so viele Änderungen vorgenommen, dass ein einzelnes Multimediaprinzip für die Beschreibungen der Änderungen gar nicht mehr ausreicht. Dennoch konnte der Fragebogen erfolgreich eingesetzt werden und die Summe der angewandten Multimediaprinzipien und deren Auswirkungen ICL, ECL und GCL abbilden.

Da die einzelnen Studien bereits in den eingebundenen Zeitschriften- und Tagungsbeiträgen diskutiert werden, wird im folgenden ein zusammenfassender Überblick über theoretische, methodische und praktische Implikationen gegeben, wobei auf einzelne Aspekte der (wünschenswerten) zukünftigen Entwicklung genauer eingegangen wird.

5.1. Theoretische Implikationen - Mehrwert

Durch die differenzierte Messung der kognitiven Belastung beim Lernen kann die Forschung die Facette der kognitiven Belastung beim Lernen genauer betrachten.

Dies ist insofern wichtig um das theoretische Konstrukt der CLT empirisch zu unterfüttern. Zudem kann eine differenzierte Messung dazu beitragen die Diskussion um das drei- oder zwei-Faktoren-Modell der CLT zu erweitern. Die Fragen sollten daher nicht lauten „Gibt es zwei oder drei Arten von kognitiver Belastung?“ oder „Wie unterscheiden sich die Arten kognitiver Belastung?“ sondern „Was muss erfasst werden, um die kognitive Belastung beim Lernen differenziert abzubilden?“. Beide Modelle haben ihre Berechtigung, die vorliegende Arbeit hat mit dem drei-Faktoren-Modell gearbeitet, da dieses besser zu Instruktionsdesignfragestellungen passt.

Wie aus Klepsch et al. (2017) jedoch hervorgeht ist gerade beim GCL nicht jedes entwickelte Item immer geeignet. Das Item mit dem Wortlaut „Die Lerneinheit enthielt Elemente, die mich unterstützten, den Lernstoff besser zu verstehen“ musste beispielsweise am Ende aus den Analysen ausgeschlossen werden. Auch in Klepsch und Seufert (2020) konnte es nicht genutzt werden, obwohl in einer Studie ein Prompt integriert war, für den dieses Item geeignet gewesen wäre. Eine mögliche Erklärung ist, dass Lernende diesen Prompt nicht als nützlich empfanden, da er im ersten Moment mehr Arbeit, die jedoch nicht automatisch als hilfreich empfunden wird, für sie bedeutet. Wer nicht an entsprechende Lernstrategien gewöhnt ist und den Prompt nicht intuitiv nutzen kann, wenn dieser also z.B. nicht nur eine Erinnerungshilfe darstellt, sondern wenn die Bearbeitung des Prompts selbst noch kognitive Ressourcen beansprucht, kann es durchaus sein, dass er nicht per se als unterstützend wahrgenommen wird, obwohl die erzeugte Belastung durch den Prompt dem GCL zuzuschreiben ist. Hier wäre Forschung über einen längeren Zeitraum nötig um festzustellen, ab wann Prompts oder sonstige geforderte Lernstrategien in Lernmaterialien nicht mehr negative kognitive Belastung auslösen, sondern ab wann sie gewinnbringend für den Lernerfolg eingesetzt werden können. In diesem Zusammenhang kann auch die Selbstregulation von Lernenden eine große Rolle spielen. Wer eigenständig (und im besten Fall erfolgreich) Lernstrategien einsetzt, kann durch andere vorgegebene Strategien ggf. weniger erfolgreich agieren (vergleiche auch: Mathematandischer Effekt; Clarck, 1990). Ein Selbstregulations-Training über einen längeren Zeitraum hinweg, kann jedoch erfolgreich Strategien festigen und die Anwendung dieser fördern (Stebner et al., 2020). Unklar bleibt noch, wie genau der Zusammenhang zwischen kognitiver Belastung und Selbstregulation aussieht. Seufert (2018) merkt an, dass eine differenzierte Messung der kognitiven Belastung dabei helfen kann herauszufinden, welche Arten kognitiver Belastung die Selbstregulation von Lernenden fördern oder behindern können, bzw. wie Selbstregulation sich auf die kognitive Belastung beim Lernen auswirkt.

Im Allgemeinen sollten die Ergebnisse durch einen systematischen Einsatz des Fragebogens in weiteren Forschungsstudien mit Multimediadesignprinzipien und Instruktionsdesignfragestellungen repliziert und genauer untersucht werden. Ich würde mir Studien wünschen die auf Basis des selben Lerninhaltes Variationen unterschiedlichster Mediendesignprinzipien überprüfen. Schön wäre auch, wenn der Inhalt in unterschiedlichen Komplexitätsstufen vorliegt und auch verschiedene Arten der kognitiven Aktivierung der Lernenden integriert werden könnten.

Des Weiteren sollten Aptitude-Treatment-Interaktionen (ATI) genauer untersucht werden. Mit Hilfe des entwickelten differenzierten Fragebogens könnte erforscht werden, wie unterschiedliche Lernermerkmale sich auf die Arten der kognitiver Belastung beim Lernen auswirken. Vielfach werden im Rahmen von Instruktionsdesignfragestellungen in Experimenten auch Lernermerkmale wie z.B. Vorwissen, räumliche Fähigkeiten oder Codalitätspräferenzen erfasst und deren Wechselwirkung mit Lernmaterialien untersucht (vergl. Brünken & Leutner, 2005). Dabei zeigte sich oftmals, dass unterschiedliche Darstellungen von Lernmaterialien auch nur für bestimmte Lerner geeignet sind. Gerade das Vorwissen, welches in der CLT als einziges Lernermerkmal in gewissem Umfang Beachtung findet, führt mitunter dazu, das bestimmte Lernmaterialien für bestimmte Lernende zu geringerer Lernleistung führen. Hierzu zählen z.B. Lernende mit zu wenig Vorwissen (vergl. z.B. Seufert, 2003) oder Lernende mit zu hohem Vorwissen (*Expertise Reversal Effekt*; vergl. z.B. Kalyuga et al., 1998). Die Nutzung des entwickelten differenzierten Fragebogens kann hier helfen, herauszufinden ob die Darstellung für Lernende mit einem bestimmten Lernermerkmal ungünstig ist und somit der ECL verändert werden sollte oder ob der Inhalt für Lernende mit bestimmten Lernermerkmalen zu komplex ist und somit erst Vorwissen aufgebaut werden muss, bevor der Inhalt in seiner aktuellen Form vermittelt werden kann. Damit kann die differenzierte Messung der kognitiven Belastung die ATI-Forschung theoriegeleitet weiterbringen und Helfen den Grund für unterschiedlichen Lernerfolg genauer beleuchten.

Weitere Forschung sollte genau hier ansetzen und Multimediadesignprinzipien und verschiedene Lernermerkmale an ein und dem selben Material in verschiedenen Varianten untersuchen. Hierfür würde sich bereits bewährtes Lernmaterial am besten eignen, welches durch verschiedene experimentelle Varianten erweitert wird.

Basierend auf den Hinweisen einiger Forscher (Seufert, 2018; Wirzberger, Herms, Esmaeili Bijarsari, Eibl & Rey, 2018; Mercimek, Akbulut, Dönmez & Sak, 2019; Mutlu-Bayraktar, Cosgun & Altan, 2019; Seufert, 2019) ist die differenzierte Messung der kognitiven Belastung erwünscht und gefordert. Wie bereits beschrieben, haben sich auch weitere Forscher bereits an einer differenzierten Messung versucht (z.B. Gerjets et al., 2006; Zumbach & Mohraz, 2008; Naismith, Cheung et al., 2015; Cierniak et al., 2009; Cierniak, 2011; Swaak & de Jong, 2001; Eysink et al., 2009; Leppink et al., 2013; Leppink, 2014). Gerade der Ansatz von Leppink et al. (2013); Leppink (2014) scheint ebenfalls erfolgversprechend, ist jedoch etwas schwerer anpassbar, da die Item-Formulierung direkt auf den Inhalt zurückgreift. Dies führt am Ende eventuell dazu, dass die unterschiedlichen Anpassungen schwer vergleichbar werden.

Einen Vergleich verschiedener Ansätze zur differenzierten Messung der kognitiven Belastung beim Lernen, wie er beispielsweise von Skulmowski und Rey (2020) erforscht wurde, sollte zukünftig öfter durchgeführt werden. Verglichen wurden drei ECL Items aus Klepsch et al. (2017) mit vier ECL Items aus Leppink et al. (2013). Skulmowski und Rey fanden, dass die Differenz zwischen den beiden Experimentalgruppen (statisches vs. interaktives Bild) in den ECL Items von Klepsch et al. (2017) größer war als in den Items von Leppink et al. (2013). Zudem konnte in

einem Haupteffekt gezeigt werden, dass sich die beiden Experimentalgruppen signifikant voneinander unterscheiden. Skulmowski und Rey (2020) konnten demnach zeigen, dass unterschiedliche Item-Formulierungen für ein und das selbe Konstrukt zu unterschiedlichen Ergebnissen führen können. Sie führen dies auf zwei Gründe zurück (Skulmowski & Rey, 2020): (1) Das Messinstrument sollte an den Lernkontext angepasst sein. Hier verweisen beide Autoren jedoch darauf, dass entsprechende Kriterien und Guidelines dazu noch fehlen. (2) Die einzelnen Arten kognitiver Belastung könnten auf multiple Belastungen je Belastungsart zurückzuführen sein (siehe auch Schnotz & Kürschner, 2007). Im vorliegenden Fall könnte der ECL aus der räumlich-visuellen Anordnung der einzelnen Elemente entstehen und/oder durch die Notwendigkeit der Interaktion selbst (z.B. Klicks ausführen).

Bereits vor einem Jahrzehnt hat de Jong (2010) Kritik an verschiedenen Aspekten der CLT geäußert. Eine differenzierte Messung der kognitiven Belastung beim Lernen kann helfen einige dieser Kritikpunkte empirisch zu untersuchen. So merkte de Jong (2010) an, dass Uneinigkeit darüber herrscht, ob ICL durch instruktionale Maßnahmen verändert werden kann. Sweller selbst vertritt hier manchmal unterschiedliche Meinungen (vergl. Sweller et al., 1998 mit van Merriënboer & Sweller, 2005). Mit Hilfe des entwickelten differenzierten Fragebogens wäre durch Replikation der Studien von z.B. Pollock, Chandler und Sweller (2002) eine Möglichkeit gegeben, herauszufinden ob beispielsweise der *Isolated Elements Effekt*, bei dem einzelne Elemente isoliert vom gesamten Lerninhalt vorab präsentiert werden, wirklich eine ICL-Variation darstellt und somit ohne Komplexitätsverlust des eigentlich zu lernenden Inhaltes eine Überlastung des Lernenden vermieden werden kann. Auch der Versuch ECL so weit wie möglich zu reduzieren wird von de Jong (2010) kritisch betrachtet: Das erstellen einer integrierten Darstellung im Vergleich zu mehreren Repräsentationen, könnte auch nach hinten los gehen, wenn die integrierte Darstellung viel zu komplex und unübersichtlich wird. Alternativ könnte das eliminieren einer redundanten Repräsentation dazu führen, dass Elaborations- und Abstraktionsprozesse verhindert werden. Ob gerade zweiteres der Fall ist, kann nur durch eine differenzierte Messung der kognitiven Belastung erforscht werden. Auch für GCL präsentiert de Jong (2010) theoretische Ungenauigkeiten, die ohne differenzierte Messung nur schwer lösbar sind und auf post-hoc Erklärungen angewiesen sind. So stellt er in den Raum, dass durch zu viel GCL wohl auch eine Überlastung möglich sein könnte. Als Beispiel wird genannt, dass eine Lernaufgabe die das Abstrahieren über mehrere konkrete Fälle erfordert, den Lernenden überfordert. Was passiert hier? Die einzelnen konkreten Fälle für sich mögen nicht komplex sein, die Darstellung könnte vollkommen in Ordnung sein und dennoch scheitert der Lernende? An was kann es liegen? Durch eine differenzierte Messung könnte genau dies festgestellt werden. Ein Beispiel: Der Lernende könnte sich zwar angestrengt und Kapazitäten in Verstehen investiert haben, aber dennoch das Lernziel nicht erreicht haben. Durch Messung der kognitive Belastung ohne Differenzierung könnte nun auch theoriekonform angenommen werden, die Darstellung des Lerninhaltes war schlecht, da die kognitiven Belastung hoch war und der Lernerfolg ausblieb. Eine differenzierte Messung könnte jedoch auch zum Schluss kommen, dass sich der Lernende zwar angestrengt hat, aber

die Aufgabenstellung so komplex war, dass er dennoch gescheitert ist. Dieses Beispiel zeigt deutlich, dass davon abgekommen werden muss, dass hoher GCL automatisch zu höherem Lernerfolg führen muss.

Die CLT fokussiert stark auf die kognitiven Aspekte des Lernens, affektive und motivationale Aspekte sind außen vor (Moreno, 2010). Das eingangs kurz erwähnte INVO-Modell (Hasselhorn & Gold, 2013) sieht jedoch genau ebensolche, in der CLT fehlenden Aspekte, auch als wichtig für gelingendes Lernen an. Mit Ausnahme der P.I.A.N.O-Studie (Rogers et al., 2014) wurden diese Aspekte zur Überprüfung des Fragebogens auch ausgeschlossen und nicht beachtet. In realen Lernumgebungen können sie natürlich nicht unbeachtet bleiben. Auch die CLT-Forschung hat dies bereits erkannt und Aspekte der lernbegleitenden Emotionen (vergl. Plass & Kalyuga, 2019; Hawthorne, Vella-Brodrick & Hattie, 2019) und Motivation (vergl. Feldon, Callan, Juth & Jeong, 2019) gingen in Überblicksarbeiten ein. Jeweils lässt sich feststellen, das die Annahme auf welche Art von kognitiver Belastung ein Aspekt wirkt noch stark hypothetisch angenommen und wenig empirisch untersucht ist. Durch die differenzierte Messung, z.B. durch den Einsatz des in der vorliegenden Dissertation entwickelten Fragebogens, besteht nun die Möglichkeit hier genauer hinzusehen und herauszufinden wie sich motivational-volitionale Aspekte auf die kognitive Belastung beim Lernen auswirken oder von diesen beeinflusst wird.

5.2. Methodische Implikationen - Stärken und Schwächen

Auch methodisch geht de Jong (2010) noch auf drei Probleme bei der Messung von kognitiver Belastung ein, die im folgenden behandelt werden: (1) Die eingesetzten Messmethoden erfassen nur die gesamte kognitive Belastung, es erfolgt keine der Theorie entsprechende Differenzierung. (2) Variationen über die Zeit finden keine Beachtung. (3) Die vorhandenen Messmethoden werden immer als relativ präsentiert: Unterschiede zwischen verschiedenen Lernmaterialien werden betrachtet, der absoluten Höhe wird keine Beachtung geschenkt. Auf diese drei Punkte wird im folgenden Abschnitt noch eingegangen, da an ihnen die Stärken und Schwächen des entwickelten Fragebogens deutlich werden.

Punkt 1 kann durch den in der vorliegende Dissertation entwickelten Fragebogen nun berücksichtigt werden, hierzu an dieser Stelle ein Rückblick auf die Entwicklung:

Die Auswahl der Items erfolgte theoriegeleitet in einer ersten Runde (siehe Studie 1 in Klepsch et al., 2017) und weniger teststatistisch, d.h. es wurde kein großer Item Pool erstellt, um daraus dann die am besten geeigneten Items abzuleiten. Da die erste Version der GCL-Items zu keinem befriedigenden Ergebnis führte wurden diese Items ausgetauscht und der Fragebogen in einer weiteren Studie erneut getestet. Im Nachhinein betrachtet muss angemerkt werden, dass die ICL-Skala lediglich aus zwei Items besteht. Ziel war jedoch auch möglichst wenig Items abzubilden. Weitere Items für ICL würden zu redundanten Formulierungen führen. Dougherty (2019) ergänzte den Fragebogen um ein weiteres Item für ICL: „I had to remember many things to perform the task.“ Ziel war eine ausbalancierte Item-Verteilung auf die Arten der

kognitiven Belastung. Mit einem Item mehr konnte auch Dougherty (2019) zeigen, dass eine Messung der kognitiven Belastung differenziert möglich ist.

Insgesamt bleibt der Fragebogen kurz genug um ihn auch öfter während einer Lerneinheit einzusetzen. Ob er sensitiv über die Zeit ist, wie von de Jong (2010) gefordert (Punkt 2), sollte in weiteren Experimenten noch überprüft werden. Objektive Methoden, die durchgängig Daten erfassen sind hier deutlich überlegen, da es schwer werden könnte, genau den passenden Zeitpunkt abzupassen um den Fragebogen zu präsentieren, zu welchem sich die kognitive Belastung während des Lernprozesses verändert. So z.B., wenn eine weitere Repräsentation ins mentale Modell integriert wird oder gerade ein Prompt gelesen wird. Dennoch ist es natürlich ein Versuch wert.

Die Praktikabilität und Effizienz von Selbstberichtsdaten macht es einfach eine große Menge von Versuchspersonendaten gleichzeitig oder sogar online zu erheben (Paulhus & Vazire, 2007). Dennoch muss angemerkt werden, dass eine subjektive Messung immer durch eine falsche Selbstwahrnehmung verzerrt werden kann (Demetriou, Ozer & Essau, 2015). Es muss jedoch auch gesagt werden, dass aktuell keine alternative objektive Messung vorliegt, die eine differenzierte Messung der Arten kognitiver Belastung beim Lernen zulässt, da keine der bekannten Methoden dazu in der Lage ist die kognitive Belastung differenziert zu erfassen. Methodisch relevant wäre daher zudem ein Abgleich von subjektiven und objektiven Methoden. Eine entsprechende Studie wurde von Korbach et al. (2017) durchgeführt: Verglichen wurden Lernmaterialien mit und ohne *Seductive Details*. Dabei konnte festgestellt werden, dass mit den objektiven Methoden Unterschiede festgestellt werden konnten, während mit den subjektiven Methoden keine Unterschiede festgestellt werden konnten. In der Studie war jedoch keine der genutzten Messungen dazu in der Lage zwischen den verschiedenen Arten kognitiver Belastung beim Lernen zu differenzieren. Die genutzte Dual-Task-Methode könnte zudem ausschlaggebend dafür sein, dass keine Unterschiede mehr in den subjektiven Skalen gefunden werden konnten. Ein Grund könnte sein, dass die Zweitaufgabe ein Teil der gesamten Aufgabe ist. In den Antworten der subjektive Messung kann diese Zweitaufgabe schwer von der Primäraufgabe getrennt werden, da sie stark mit dieser verwoben ist.

Dennoch sollten ähnliche Studien mit weniger ablenkenden Messmethoden durchgeführt werden um ein umfassenderes Bild über das Zusammenspiel von subjektiven und objektiven Methoden zu erhalten. Viele Arten der objektiven Messung werden bei einer Differenzierung der kognitiven Belastung beim Lernen wenig hilfreich sein. Eye-Tracking jedoch könnte Möglichkeiten in diese Richtung aufweisen. Beispielsweise könnte weitergehende Forschung versuchen herauszufinden, inwieweit die verschiedenen erfassten Parameter beim Eye-Tracking unterschiedlichen Arten der kognitiven Belastung beim Lernen zugeordnet werden können. Betrachtet werden beim Eye-Tracking meist Fixationen (Punkte, die man genau betrachtet), Sakkaden (schnellen Augenbewegungen) und Regressionen (Rücksprünge). Gerade Regressionen könnten auf Aktivitäten hindeuten, die versuchen tieferes Verständnis zu generieren (Eskenazi & Folk, 2017).

Bleibt noch der dritte Kritikpunkt zur Messung kognitiver Belastung von de Jong

(2010) offen: Es wird meist nur der Unterschied in der Höhe der kognitiven Belastung zwischen zwei unterschiedlichen Lerneinheiten betrachtet. Auch die vorliegende differenzierte Messung der kognitiven Belastung beim Lernen ist für diesen Einsatz ausgelegt und auch die in den unterschiedlichen Beiträgen präsentierten Studien machen genau dies. Da die kognitive Belastung stark von Lernereigenschaften (dazu im nächsten Abschnitt mehr) abhängig ist, ist es nicht möglich, die zu erwartende kognitive Belastung für ein bestimmtes Lernmaterial für alle Lernenden auf eine bestimmte „Menge“ auf einer gegebenen Skala festzumachen. Dieser Kritikpunkt wird methodisch somit auch über die vorliegende differenzierte Messung nicht lösbar sein.

5.3. Praktische Implikationen - Ausblick

Die acht Fragen des differenzierten CLT-Fragebogens, wie er in Klepsch et al. (2017) veröffentlicht wurde, sind zusammen in knapp fünf Minuten gut zu beantworten. Damit wird es möglich in kurzer Zeit die kognitive Belastung beim Lernen differenziert zu messen. Durch die Kürze können die Items schnell präsentiert, überall eingebunden und von den Versuchspersonen oder Lernenden auch rasch beantwortet werden, ohne dass dafür viel Zeit eingeplant werden muss. Wie von van Gog et al. (2012) festgestellt, macht es einen Unterschied ob mehrfach nach Teilaufgaben oder einmalig nach der Gesamtaufgabe die kognitive Belastung erfasst wird. In 4 Experimenten haben van Gog et al. (2012) das Item von Paas (1992) genutzt und konnten zeigen, dass die mentale Belastung bei einer Abfrage am Ende als höher eingeschätzt wird als das Mittel über mehrere Abfragen während der Aufgabenbearbeitung. Ein ähnliches Experiment wäre für die entwickelte differenzierte Messung noch sinnvoll. Der differenzierte Fragebogen ist kurz genug um den Lernprozess nicht unnötig zu stören, ein entsprechendes Experiment könnte zeigen ob sich die Arten kognitiver Belastung über Teilaufgaben hinweg verändern. So könnte für Designeffekte ein Gewöhnungseffekt eintreten und schlechtes Design über mehrere Teilaufgaben hinweg weniger relevant werden, da Lernende irrelevante Designeffekte ausblenden können. Dies könnte z.B. für das *Seductive Details Prinzip* (Rey, 2012; Sundararajan & Adesope, 2020) relevant sein. Lernende könnten nach Bearbeitung der ersten Teilaufgaben bereits feststellen, dass z.B. bestimmte Abbildungen irrelevant für ihren Lernerfolg sind und können diese somit in späteren Teilaufgaben schneller als irrelevant erkennen. Alternativ könnte die Elementinteraktivität über die Teilaufgaben hinweg abnehmen, da das mentale Modell mit jeder Teilaufgabe bereits weiter angereichert wurde. Beide Beispiele würden dazu führen, dass die kognitive Belastung über die Zeit abnehmen würde, der Grund jedoch ein völlig anderer ist. Die differenzierte Messung der kognitiven Belastung könnte hier genauere Einblicke ins Lernen geben.

Da der Fragebogen, wie eben schon erklärt, sehr kurz ist, können auch Lehrende und Instruktionsdesigner die kognitive Belastung des Lerninhaltes bzw. Lernmaterials erfassen und feststellen wo Lernschwierigkeiten bestehen und somit auch wo nachgearbeitet werden muss.

Beim Vergleich mehrerer Lernmaterialien kann gerade die ICL-Skala auch eingesetzt werden um festzustellen, ob die Komplexität des Lerninhaltes durch eine Darstellungsänderung entscheidend verändert wurde. Genau zu diesem Zweck nutzten Eitel, Bender und Renkl (2019) die ICL-Skala: In einer Studie zum *Seductive Details Effekt* wurde die ICL-Skala als Kontrollvariable eingesetzt. Somit konnte überprüft werden, ob die intrinsische Belastung in den drei Experimentalgruppen durch das Hinzufügen von Seductive Details verändert wurde (was im übrigen nicht der Fall war). Ein Einsatz ähnlich wie bei Eitel et al. (2019) ist begrüßenswert um festzustellen ob Designveränderungen (wenn auch versehentlich) Auswirkungen auf die Elementinteraktivität haben. Von Sweller (2010) wurde argumentiert, dass Elementinteraktivität auf alle Arten kognitiver Belastung Auswirkungen haben kann: Wenn die Elementinteraktivität verändert werden kann ohne zu ändern was (inhaltlich) gelernt wurde, so liegt eine Veränderung der extrinsischen Belastung vor. Liegt jedoch einer Veränderung des Inhalts vor, so führt die dadurch veränderte Elementinteraktivität zu einer veränderten intrinsischen Belastung (siehe auch Beckmann, 2010). Durch die Nutzung der ICL-Skala zur Gegenprüfung ob eine Designveränderung auch ungewollt eine Veränderung der inhaltlichen Komplexität des Lerninhaltes nach sich zieht, kann daher die ICL-Skala dazu genutzt werden, nachzuweisen dass das Lernmaterial der unterschiedlichen Experimentalgruppen sich nicht in der inhaltlichen Komplexität unterscheidet. Zudem konnte der Fragebogen bereits in Studien erfolgreich eingesetzt werden (z.B. Lehmann & Seufert, 2017; Eitel et al., 2019; Dougherty, 2019; Lehmann & Seufert, 2019; Mille et al., 2019). Die differenzierte Messung stellt demnach ein wertvolles Tool zum Verstehen von Lernprozessen dar.

Außerdem kann, wie in Rogers et al. (2014) gezeigt, in Kombination mit weiteren Erhebungsinstrumenten gerade in komplexen und/oder digitalen Lernumgebungen ein umfassenderes Bild des Lernenden aufgezeigt werden.

5.4. Abschlussbemerkung

Die differenzierte Messung kognitiver Belastung mit Hilfe eines subjektiven Fragebogens kann helfen Lernprozesse zu verstehen und kann vor allem dazu genutzt werden in realen Lernsituationen als Evaluation des bereitgestellten Lernmaterials zu dienen. Dabei ist gerade die kürze des Fragebogens ausschlaggebend für den praktischen Einsatz. Weitere Forschung ist nötig, aber kein Hindernis, den Fragebogen bereits jetzt - als Instruktionsdesigner/-in oder Lehrende/-r - für die eigene Auswertung von Materialien zu nutzen um ein Gefühl für die kognitive Belastung der eigenen Lernenden zu erhalten.

Durch die durchgeführten Studien kann zwar nicht abschließend festgestellt werden ob das drei- oder zwei-Faktoren-Modell der CLT zu bevorzugen ist, es konnte jedoch hoffentlich deutlich gemacht werden, dass die praktische Relevanz der nötigen Messung aller für Instruktionsdesignfragestellungen relevanter Aspekte der kognitiven Belastung in der Forschung im Vordergrund stehen sollte. Dennoch trägt diese

Arbeit natürlich auch zur Theoriebildung bei, indem sie sich auf die Seite der Verfechter eines drei-Faktoren-Modells stellt. Insbesondere gesteht das drei-Faktoren-Modell Lehrenden einen Einfluss auf den GCL zu. Laut Paas und van Gog (2006) starten Lernende selten spontan Aktivitäten welche die Schemakonstruktion fördern, zu tieferem Verständnis führen und den Transfer des gelernten auf andere Situationen erlauben. Aus diesem Grund sind Maßnahmen zur Förderung des GCL durch Lehrende durchaus wichtig und ihre Effektivität sollte auch entsprechend - z.B. durch eine differenzierte Messung der kognitiven Belastung - erforscht werden können.

Ich zumindest wünsche mir für meine drei angehenden Schüler - Julius, Anastasia und Sebastian -, dass ihre Lehrer und Lehrerinnen wissen und erfassen können ob ihr Einsatz und die Arbeit die sie investieren effektiver ist als wenn den Schülern „nur“ ein Buch vorgelegt wird ... Vielleicht sollte auch erforscht werden, welche Auswirkung der Lehrende selbst mit seiner Anwesenheit auf die Lernleistung und kognitive Belastung beim Bearbeiten von Lernmaterial hat. Gerade für die Online-Lehre könnte auch die Präsenz des Lehrenden einen relevanten Unterschied ausmachen, da Präsenz (hoffentlich) auch mit Instruktion, Fokussierung auf Relevantes und der Darbietung von Hilfestellung einhergeht.

Literaturverzeichnis

Aldekhyl, S., Cavalcanti, R. B. & Naismith, L. M. (2018). Cognitive load predicts point-of-care ultrasound simulator performance. *Perspectives on medical education*, *7* (1), 23–32. doi: 10.1007/s40037-017-0392-7

Atkinson, R. K. & Shiffrin, R. M. (1971). *The Control Processes of Short-Term Memory* (Nr. 173). Stanford, California.

Ayres, P. (2006). Impact of reducing intrinsic cognitive load on learning in a mathematical domain. *Applied Cognitive Psychology*, *20* (3), 287–298. doi: 10.1002/acp.1245

Ayres, P. (2011). Rethinking Germane Cognitive Load. In European Association for Research on Learning and Instruction (Hrsg.), *EARLI Conference 2011 "Education for a Global Networked Society"* (S. 1768–1770). Exeter: University of Exeter.

Baddeley, A. D. (1986). *Working memory.* Oxford, New York: Oxford University Press.

Baddeley, A. D. (1994). The magical number seven: Still magical after all these years? *Psychological Review*, *101* (2), 353–356.

Bannert, M. (2002). Managing cognitive load - Recent trends in cognitive load theory. *Learning and Instruction*, *12* (1), 139–146. doi: 10.1016/S0959-4752(01)00021-4

Bannert, M. (2009). Promoting Self-Regulated Learning Through Prompts. *Zeitschrift für Pädagogische Psychologie*, *23* (2), 139–145. doi: 10.1024/1010-0652.23.2.139

Beckmann, J. F. (2010). Taming a beast of burden – On some issues with the conceptualisation and operationalisation of cognitive load. *Learning and Instruction*, *20* (3), 250–264. doi: 10.1016/j.learninstruc.2009.02.024

Bisra, K., Liu, Q., Nesbit, J. C., Salimi, F. & Winne, P. H. (2018). Inducing Self-Explanation: a Meta-Analysis. *Educational Psychology Review*, *30* (3), 703–725. doi: 10.1007/s10648-018-9434-x

Brünken, R. & Leutner, D. (2005). Individuelle Unterschiede beim Lernen mit neuen Medien – neue Wege in der ATI-Forschung? In S. R. Schilling, J. R. Sparfeldt & C. Pruisken (Hrsg.), *Aktuelle Aspekte pädagogisch-psychologischer Forschung.* Münster: Waxmann.

Brünken, R., Plass, J. L. & Leutner, D. (2003). Direct Measurement of Cognitive Load in Multimedia Learning. *Educational Psychologist*, *38* (1), 53–61.

Brünken, R., Plass, J. L. & Leutner, D. (2004). Assessment of Cognitive Load in Multimedia Learning with Dual-Task Methodology: Auditory

Load and Modality Effects. *Instructional Science*, *32* (1), 115–132. doi: 10.1023/B:TRUC.0000021812.96911.c5

Brünken, R., Seufert, T. & Paas, F. G. W. C. (2010). Measuring Cognitive Load. In J. L. Plass, R. Moreno & R. Brünken (Hrsg.), *Cognitive Load Theory* (S. 181–202). Cambridge, New York: Cambridge University Press.

Brünken, R., Steinbacher, S., Plass, J. L. & Leutner, D. (2002). Assessment of Cognitive Load in Multimedia Learning Using Dual-Task Methodology. *Experimental Psychology*, *49* (2), 109–119. doi: 10.1027//1618-3169.49.2.109

Butcher, K. R. (2014). The Multimedia Principle. In R. E. Mayer (Hrsg.), *The Cambridge handbook of multimedia learning* (S. 174–205). New York: Cambridge University Press. doi: 10.1017/CBO9781139547369.010

Chandler, P. & Sweller, J. (1991). Cognitive Load Theory and the Format of Instruction. *Cognition & Instruction*, *8* (4), 293–332.

Chen, F., Zhou, J., Wang, Y., Yu, K., Arshad, S. Z., Khawaji, A. & Conway, D. (2016). *Robust Multimodal Cognitive Load Measurement*. Cham and s.l.: Springer International Publishing. Zugriff auf `http://gbv.eblib.com/patron/FullRecord.aspx?p=4557252` doi: 10.1007/978-3-319-31700-7

Cierniak, G. (2011). *Facilitating and Inhibiting Learning by the Spatial Contiguity of Text and Graphic: How Does Cognitive Load Mediate the Split-Attention and Expertise Reversal Effect?* (Dissertation). Eberhard Karls Universität Tübingen, Tübingen.

Cierniak, G., Scheiter, K. & Gerjets, P. (2009). Explaining the split-attention effect: Is the reduction of extraneous cognitive load accompanied by an increase in germane cognitive load? *Computers in Human Behavior*, *25* (2), 315–324. doi: 10.1016/j.chb.2008.12.020

Clarck, R. E. (1990). When teaching kills learning: research on mathemathantics. In H. Mandl, E. de Corte, N. Bennett & H. F. Friedrich (Hrsg.), *Analysis of complex skills and complex knowledge domains: Selection of papers from the Second European Conference for Research on Learning and Instruction held in Tübingen, West Germany, in September 1987* (S. 1–22). Oxford: Pergamon Press.

Clark, R. C., Nguyen, F. & Sweller, J. (2006). *Efficiency in learning: Evidence-based guidelines to manage cognitive load*. San Francisco: Pfeiffer.

Cooper, G. & Sweller, J. (1987). Effects of Schema Acquisition and Rule Automation on Mathematical Problem-Solving Transfer. *Journal of Educational Psychology*, *79* (4), 347–362.

Cowan, N. (1995). *Attention and memory: An integrated framework* (Bd. 26). New York: Oxford Univ. Press.

Cowan, N. (2001). The magical number 4 in short-term memory: A reconsideration of mental storage capacity. *Behavioral and Brain Sciences*, *24*, 87–114.

Cowan, N. (2007). An Embedded Process Model of Working Memory. In A. Miyake & P. Shah (Hrsg.), *Models of working memory* (S. 62–101). Cambridge: Cambridge Univ. Press.

Cowan, N. (2010). The Magical Mystery Four: How is Working Memory Capacity Limited, and Why? *Current directions in psychological science*, *19* (1), 51–57. doi: 10.1177/0963721409359277

de Jong, T. (2010). Cognitive load theory, educational research, and instructional design: some food for thought. *Instructional Science*, *38* (2), 105–134. doi: 10.1007/s11251-009-9110-0

DeLeeuw, K. E. & Mayer, R. E. (2008). A comparison of three measures of cognitive load: Evidence for separable measures of intrinsic, extraneous, and germane load. *Journal of Educational Psychology*, *100* (1), 223–234. doi: 10.1037/0022-0663.100.1.223

Demetriou, C., Ozer, B. U. & Essau, C. A. (2015). Self-Report Questionnaires. In R. L. Cautin & S. O. Lilienfeld (Hrsg.), *The encyclopedia of clinical psychology* (S. 1–6). doi: 10.1002/9781118625392.wbecp507

Diemand-Yauman, C., Oppenheimer, D. M. & Vaughan, E. B. (2011). Fortune favors the Bold (and the Italicized): Effects of disfluency on educational outcomes. *Cognition*, *118* (1), 111–115. doi: 10.1016/j.cognition.2010.09.012

Dougherty, S. (2019). *Partnering People with Deep Learning Systems: Human Cognitive Effects of Explanations* (Dissertation). Georgia State University, Atlanta.

Duchowski, A. T., Krejtz, K., Krejtz, I., Biele, C., Niedzielska, A., Kiefer, P., … Giannopoulos, I. (2018). The Index of Pupillary Activity. In R. Mandryk, M. Hancock, M. Perry & A. Cox (Hrsg.), *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems - CHI '18* (S. 1–13). New York, New York, USA: ACM Press. doi: 10.1145/3173574.3173856

Eitel, A., Bender, L. & Renkl, A. (2019). Are seductive details seductive only when you think they are relevant? An experimental test of the moderating role of perceived relevance. *Applied Cognitive Psychology*, *33* (1), 20–30. doi: 10.1002/acp.3479

Eskenazi, M. A. & Folk, J. R. (2017). Regressions during reading: The cost depends on the cause. *Psychonomic Bulletin & Review*, *24* (4), 1211–1216. doi: 10.3758/s13423-016-1200-9

Eysink, T. H. S., de Jong, T., Berthold, K., Kolloffel, B., Opfermann, M. & Wouters, P. (2009). Learner Performance in Multimedia Learning Arrangements: An Analysis Across Instructional Approaches. *American Educational Research Journal*, *46* (4), 1107–1149. doi: 10.3102/0002831209340235

Feldon, D. F., Callan, G., Juth, S. & Jeong, S. (2019). Cognitive Load as Motivational Cost. *Educational Psychology Review*, *31* (2), 319–337. doi: 10.1007/s10648-019-09464-6

Florax, M. & Ploetzner, R. (2010). What contributes to the split-attention effect? The role of text segmentation, picture labelling, and spatial proximity. *Learning and Instruction*, *20* (3), 216–224. doi: 10.1016/j.learninstruc.2009.02.021

Gerjets, P., Scheiter, K. & Catrambone, R. (2006). Can learning from molar and modular worked examples be enhanced by providing instructional explanations and prompting self-explanations? *Learning and Instruction*, *16* (2), 104–121.

doi: 10.1016/j.learninstruc.2006.02.007

Ginns, P. (2005). Meta-analysis of the modality effect. *Learning and Instruction*, *15* (4), 313–331. doi: 10.1016/j.learninstruc.2005.07.001

Gopher, D. & Braune, R. (1984). On the Psychophysics of Workload: Why Bother with Subjective Measures? *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *26* (5), 519–532. doi: 10.1177/001872088402600504

Hart, S. G. & Staveland, L. E. (1988). Development of NASA-TLX (Task Load Index): Results of Empirical and Theoretical Research. In N. Meshkati & P. A. Hancock (Hrsg.), *Human Mental Workload* (S. 139–183). North-Holland: Elsevier.

Hasselhorn, M. & Gold, A. (2013). *Pädagogische Psychologie: Erfolgreiches Lernen und Lehren* (3., vollständig überarbeitete und erweiterte Auflage Aufl.). Stuttgart: Verlag W. Kohlhammer. Zugriff auf `http://gbv.eblib.com/patron/FullRecord.aspx?p=1766654`

Hasselhorn, M. & Gold, A. (2017). *Pädagogische Psychologie: Erfolgreiches Lernen und Lehren* (4., aktualisierte Auflage Aufl.). Stuttgart: Verlag W. Kohlhammer.

Hawthorne, B. S., Vella-Brodrick, D. A. & Hattie, J. (2019). Well-Being as a Cognitive Load Reducing Agent: A Review of the Literature. *Frontiers in Education*, *4*. doi: 10.3389/feduc.2019.00121

Kalyuga, S. (2011). Cognitive Load Theory: How Many Types of Load Does It Really Need? *Educational Psychology Review*, *23* (1), 1–19. doi: 10.1007/s10648-010-9150-7

Kalyuga, S., Chandler, P. & Sweller, J. (1998). Levels of Expertise and Instructional Design. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *40* (1), 1–17. doi: 10.1518/001872098779480587

Kester, L., Kirschner, P. A. & van Merriënboer, J. J. (2004). Timing of Information Presentation in Learning Statistics. *Instructional Science*, *32* (3), 233–252. doi: 10.1023/B:TRUC.0000024191.27560.e3

Kirschner, P. A., Ayres, P. & Chandler, P. (2011). Contemporary cognitive load theory research: The good, the bad and the ugly. *Computers in Human Behavior*, *27* (1), 99–105. doi: 10.1016/j.chb.2010.06.025

Klepsch, M., Kempter, I. & Seufert, T. (2013a). Interaction of Worked Examples and Prompts: Impact on Performance and Cognitive Load. In European Association for Research on Learning and Instruction (Hrsg.), *Responsible Teaching and Sustainable Learning* (S. 983).

Klepsch, M., Kempter, I. & Seufert, T. (2013b). Interaction of Worked Examples and Prompts: Impact on Performance and Cognitive Load. In European Association for Research on Learning and Instruction (Hrsg.), *Responsible Teaching and Sustainable Learning* (S. 983).

Klepsch, M., Könings, B., Weber, M. & Seufert, T. (25.-27.08.2014). *Fostering Piano Learning by dynamic mapping of notes: Paper.* Rotterdam (NE).

Klepsch, M., Schmitz, F. & Seufert, T. (2017). Development and Validation of Two Instruments Measuring Intrinsic, Extraneous, and Germane Cognitive Load.

Frontiers in Psychology, *8*, 1997. doi: 10.3389/fpsyg.2017.01997

Klepsch, M. & Seufert, T. (09.-11.04.2012). *Subjective Differentiated Measurement of Cognitive Load: Paper.* Tallahassee (USA).

Klepsch, M. & Seufert, T. (2020). Understanding instructional design effects by differentiated measurement of intrinsic, extraneous, and germane cognitive load. *Instructional Science*, *48* (1), 45–77. doi: 10.1007/s11251-020-09502-9

Klepsch, M., Westphal, J. & Seufert, T. (2013). Effekte von integriertem bzw. separiertem Text-Bild-Material in Abhängigkeit von serialistischem oder holistischem Lernstil: Paper. In Fachgruppe Pädagogische Psychologie (Hrsg.), *Gemeinsam verschieden* (S. 31). Hildesheim (DE).

Korbach, A., Brünken, R. & Park, B. (2017). Measurement of cognitive load in multimedia learning: a comparison of different objective measures. *Instructional Science*, *45* (4), 515–536. doi: 10.1007/s11251-017-9413-5

Lavie, N. (1995). Perceptual load as a necessary condition for selective attention. *Journal of Experimental Psychology: Human Perception and Performance*, *21* (3), 451–468. doi: 10.1037/0096-1523.21.3.451

Lehmann, J. A. M. & Seufert, T. (2017). The Influence of Background Music on Learning in the Light of Different Theoretical Perspectives and the Role of Working Memory Capacity. *Frontiers in Psychology*, *8*, 1902. doi: 10.3389/fpsyg.2017.01902

Lehmann, J. A. M. & Seufert, T. (2019). The Interaction Between Text Modality and the Learner's Modality Preference Influences Comprehension and Cognitive Load. *Frontiers in Psychology*, *10*, 2820. doi: 10.3389/fpsyg.2019.02820

Leppink, J. (2014). Managing the load on a reader's mind. *Perspectives on medical education*, *3* (5), 327–328. doi: 10.1007/s40037-014-0144-x

Leppink, J., Paas, F. G. W. C., van der Vleuten, C. P. M., van Gog, T. & van Merriënboer, J. J. G. (2013). Development of an instrument for measuring different types of cognitive load. *Behavior research methods*, *45* (4), 1058–1072. doi: 10.3758/s13428-013-0334-1

Leppink, J., Paas, F. G. W. C., van Gog, T., van der Vleuten, C. P. M. & van Merriënboer, J. J. G. (2014). Effects of pairs of problems and examples on task performance and different types of cognitive load. *Learning and Instruction*, *30*, 32–42. doi: 10.1016/j.learninstruc.2013.12.001

Mawer, R. F. & Sweller, J. (1982). Effects of subgoal density and location on learning during problem solving. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *8* (3), 252–259. doi: 10.1037/0278-7393.8.3.252

Mayer, R. E. (2009). *Multimedia learning* (2. Aufl.). Cambridge: Cambridge University Press.

Mercimek, B., Akbulut, Y., Dönmez, O. & Sak, U. (2019). Multitasking impairs learning from multimedia across gifted and non-gifted students. *Educational Technology Research and Development*. doi: 10.1007/s11423-019-09717-9

Mietzel, G. (2017). *Pädagogische Psychologie des Lernens und Lehrens* (9., aktualisierte und erweiterte Auflage Aufl.). Göttingen: Hogrefe. Zugriff auf `https://elibrary.hogrefe.de/9783840924576/` doi: 10.1026/02457-000

Mille, C., Fleury, S., Pasi, S., Fournier, K., Izzouzi, L., Duchossoy, S., ... Richir, S. (2019). *Effets de stimuli externes non pertinents sur la créativité.* Lyon. Zugriff auf `https://hal.archives-ouvertes.fr/hal-02303945`

Miller, G. A. (1956a). Information and Memory. *Scientific American*, *195* (2), 42–46. doi: 10.1038/scientificamerican0856-42

Miller, G. A. (1956b). The magical number seven, plus or minus two: some limits on our capacity for processing information. *Psychological Review*, *63* (2), 81–97. doi: 10.1037/h0043158

Miller, G. A. (1989). George A. Miller. In G. Lindzey (Hrsg.), *A history of psychology in autobiography Vol. VIII* (S. 390–418). Stanford University Press. doi: 10.1037/11347-011

Miyake, A. & Shah, P. (Hrsg.). (2007). *Models of working memory: Mechanisms of active maintenance and executive control* (Reprint Aufl.). Cambridge: Cambridge Univ. Press.

Moreno, R. (2005). Instructional Technology: Promise and Pitfalls. In L. M. PytlikZillig, M. Bodvarsson & R. Bruning (Hrsg.), *Technology-based education* (S. 1–19). Greenwich, Conn.: Information Age Pub.

Moreno, R. (2010). Cognitive load theory: more food for thought. *Instructional Science*, *38* (2), 135–141. doi: 10.1007/s11251-009-9122-9

Moreno, R. & Park, B. (2010). Cognitive Load Theory: Historical Development and Relation to Other Theories. In J. L. Plass, R. Moreno & R. Brünken (Hrsg.), *Cognitive Load Theory* (S. 9–28). Cambridge, New York: Cambridge University Press.

Müller, H. & Krummenacher, J. (2012). Funktionen und Modelle der selektiven Aufmerksamkeit. In H.-O. Karnath & P. Thier (Hrsg.), *Kognitive Neurowissenschaften* (S. 307–321). Berlin, Heidelberg: Springer Berlin Heidelberg. doi: 10.1007/978-3-642-25527-4_28

Mutlu-Bayraktar, D., Cosgun, V. & Altan, T. (2019). Cognitive load in multimedia learning environments: A systematic review. *Computers & Education*, *141*, 103618. doi: 10.1016/j.compedu.2019.103618

Naismith, L. M., Cheung, J. J. H., Ringsted, C. & Cavalcanti, R. B. (2015). Limitations of subjective cognitive load measures in simulation-based procedural training. *Medical Education*, *49* (8), 805–814. doi: 10.1111/medu.12732

Naismith, L. M., Haji, F. A., Sibbald, M., Cheung, J. J. H., Tavares, W. & Cavalcanti, R. B. (2015). Practising what we preach: using cognitive load theory for workshop design and evaluation. *Perspectives on medical education*, *4* (6), 344–348. doi: 10.1007/s40037-015-0221-9

NSW Department of Education. (2017). *Cognitive load theory: Research that teachers really need to understand.* Sydney.

Okuni, I. M. & Widyanti, A. (2019). International Students' Cognitive Load in Learning through a Foreign Language of Instruction: A Case of Learning using Bahasa - Indonesia. *PEOPLE: International Journal of Social Sciences*, *4* (3), 1503–1532. doi: 10.20319/pijss.2019.43.15031532

Owen, E. & Sweller, J. (1985). What do students learn while solving mathema-

tics problems? *Journal of Educational Psychology*, *77* (3), 272–284. doi: 10.1037/0022-0663.77.3.272

Paas, F. G. W. C. (1992). Training Strategies for Attaining Transfer of Problem-Solving Skill in Statistics: A Cognitive-Load Approach. *Journal of Educational Psychology*, *84* (4), 429–434.

Paas, F. G. W. C., Renkl, A. & Sweller, J. (2003). Cognitive Load Theory and Instructional Design: Recent Developments. *Educational Psychologist*, *38* (1), 1–4. doi: 10.1207/S15326985EP3801_1

Paas, F. G. W. C., Tuovinen, J. E., Tabbers, H. K. & van Gerven, P. W. M. (2003). Cognitive Load Measurement as a Means to Advance Cognitive Load Theory. *Educational Psychologist*, *38* (1), 63–71.

Paas, F. G. W. C. & van Gog, T. (2006). Optimising worked example instruction: Different ways to increase germane cognitive load. *Learning and Instruction*, *16* (2), 87–91. doi: 10.1016/j.learninstruc.2006.02.004

Paas, F. G. W. C., van Merriënboer, J. J. & Adam, J. J. (1994). Measurement of cognitive load in instructional research. *Perceptual and Motor Skills*, *79* (1 Pt 2), 419–430. doi: 10.2466/pms.1994.79.1.419

Paivio, A. (1971). *Imagery and verbal processes.* New York: Holt Rinehart and Winston.

Paivio, A. (1986). *Mental Representations: A Dual Coding Approach* (Bd. 9). Oxford: Oxford University Press.

Park, B. (2010). *Testing the Additivity Hypothesis of Cognitive Load Theory* (Dissertation). Universität des Saarlandes, Saarbrücken.

Park, B. & Brünken, R. (2015). The Rhythm Method: A New Method for Measuring Cognitive Load-An Experimental Dual-Task Study. *Applied Cognitive Psychology*, *29* (2), 232–243. doi: 10.1002/acp.3100

Paulhus, D. L. & Vazire, S. (2007). The self-report method. In R. W. Robins, R. C. Fraley & R. F. Krueger (Hrsg.), *Handbook of research methods in personality psychology* (S. 224–239). New York, NY, US: The Guilford Press.

Pichler, M. & Seufert, T. (2011). Two strategies to measure Cognitive Load. In European Association for Research on Learning and Instruction (Hrsg.), *EARLI Conference 2011 "Education for a Global Networked Society"* (S. 928–929). Exeter: University of Exeter.

Plass, J. L. & Kalyuga, S. (2019). Four Ways of Considering Emotion in Cognitive Load Theory. *Educational Psychology Review*, *31* (2), 339–359. doi: 10.1007/s10648-019-09473-5

Pollock, E., Chandler, P. & Sweller, J. (2002). Assimilating complex information. *Learning and Instruction*, *12* (1), 61–86. doi: 10.1016/S0959-4752(01)00016-0

Rahimi, M. & Sayyadi, M. (2019). The cognitive load of listening activities of a cognitive-based listening instruction. *Indonesian Journal of Applied Linguistics*, *9* (2). doi: 10.17509/ijal.v9i2.20236

Renkl, A. (2015). Wissenserwerb. In E. Wild & J. Möller (Hrsg.), *Pädagogische Psychologie* (S. 3–24). Berlin, Heidelberg and s.l.: Springer Berlin Heidelberg.

Renkl, A. & Scheiter, K. (2017). Studying Visual Displays: How to Instructionally Support Learning. *Educational Psychology Review*, *29* (3), 599–621. doi: 10.1007/s10648-015-9340-4

Rey, G. D. (2009). *E-Learning: Theorien, Gestaltungsempfehlungen und Forschung* (1. Auflage Aufl.). Bern: Verlag Hans Huber. Zugriff auf `http://www.socialnet.de/rezensionen/isbn.php?isbn=978-3-456-84743-6`

Rey, G. D. (2012). A review of research and a meta-analysis of the seductive detail effect. *Educational Research Review*, *7* (3), 216–237. doi: 10.1016/j.edurev.2012.05.003

Rogers, K., Röhlig, A., Weing, M., Gugenheimer, J., Könings, B., Klepsch, M., ... Weber, M. (2014). P.I.A.N.O. In R. Dachselt, N. Graham, K. Hornbæk & M. Nacenta (Hrsg.), *ITS '14 Proceedings of the Ninth ACM International Conference on Interactive Tabletops and Surfaces* (S. 149–158). doi: 10.1145/2669485.2669514

Salomon, G. (1984). Television is "easy" and print is "tough": The differential investment of mental effort in learning as a function of perceptions and attributions. *Journal of Educational Psychology*, *76* (4), 647–658. doi: 10.1037/0022-0663.76.4.647

Saults, J. S. & Cowan, N. (2007). A central capacity limit to the simultaneous storage of visual and auditory arrays in working memory. *Journal of experimental psychology. General*, *136* (4), 663–684. doi: 10.1037/0096-3445.136.4.663

Schnotz, W. & Kürschner, C. (2007). A Reconsideration of Cognitive Load Theory. *Educational Psychology Review*, *19* (4), 469–508. doi: 10.1007/s10648-007-9053-4

Schroeder, N. L. & Cenkci, A. T. (2018). Spatial Contiguity and Spatial Split-Attention Effects in Multimedia Learning Environments: a Meta-Analysis. *Educational Psychology Review*, *30* (3), 679–701. doi: 10.1007/s10648-018-9435-9

Seufert, T. (2003). Supporting coherence formation in learning from multiple representations. *Learning and Instruction*, *13* (2), 227–237. doi: 10.1016/S0959-4752(02)00022-1

Seufert, T. (2018). The interplay between self-regulation in learning and cognitive load. *Educational Research Review*, *24*, 116–129. doi: 10.1016/j.edurev.2018.03.004

Seufert, T. (2019). Training for Coherence Formation When Learning From Text and Picture and the Interplay With Learners' Prior Knowledge. *Frontiers in Psychology*, *10*, 193. doi: 10.3389/fpsyg.2019.00193

Seufert, T., Klepsch, M. & Westphal, J. (2017). Adjusting Task Difficulty to Control Learner's Intrinsic Cognitive Load. In European Association for Research on Learning and Instruction (Hrsg.), *EARLI 2017 Book of Abstracts* (S. 682).

Skulmowski, A. & Rey, G. D. (2020). Subjective cognitive load surveys lead to divergent results for interactive learning media. *Human Behavior and Emerging Technologies*. doi: 10.1002/hbe2.184

Smallwood, J. & Schooler, J. W. (2014). Mind-Wandering. In T. Bayne, A. Cleeremans & P. Wilken (Hrsg.), *The Oxford companion to consciousness* (S. 443–445). Oxford: Oxford Univ. Press.

Stebner, F., Schuster, C., Dicke, T., Karlen, Y., Wirth, J. & Leutner, D. (2020). The effects of self-regulation training on self-regulated learning competencies and cognitive load: Does socioeconomic status matter? In S. Tindall-Ford, S. Agostinho & J. Sweller (Hrsg.), *Advances in cognitive load theory* (S. 194–208). Milton Park, Abingdon, Oxon and New York, NY: Routledge.

Sundararajan, N. & Adesope, O. (2020). Keep it Coherent: A Meta-Analysis of the Seductive Details Effect. *Educational Psychology Review*. doi: 10.1007/s10648-020-09522-4

Swaak, J. & de Jong, T. (2001). Learner vs. System Control in Using Online Support for Simulation-based Discovery Learning. *Learning Environments Research*, *4* (3), 217–241. doi: 10.1023/A:1014434804876

Sweller, J. (1983). Control mechanisms in problem solving. *Memory & Cognition*, *11* (1), 32–40. doi: 10.3758/BF03197659

Sweller, J. (2005). Implications of Cognitive Load Theory for Multimedia Learning. In R. E. Mayer (Hrsg.), *The Cambridge handbook of multimedia learning* (S. 19–30). Cambridge, New York: Cambridge University Press.

Sweller, J. (2010). Element Interactivity and Intrinsic, Extraneous, and Germane Cognitive Load. *Educational Psychology Review*, *22* (2), 123–138. doi: 10.1007/s10648-010-9128-5

Sweller, J., Ayres, P. & Kalyuga, S. (2011). *Cognitive load theory*. New York: Springer.

Sweller, J. & Chandler, P. (1994). Why Some Material Is Difficult to Learn. *Cognition & Instruction*, *12* (3), 185.

Sweller, J. & Cooper, G. A. (1985). The Use of Worked Examples as a Substitute for Problem Solving in Learning Algebra. *Cognition & Instruction*, *2* (1), 59–89. doi: 10.1207/s1532690xci0201_3

Sweller, J., Mawer, R. F. & Howe, W. (1982). Consequences of History-Cued and Means-End Strategies in Problem Solving. *The American Journal of Psychology*, *95* (3), 455. doi: 10.2307/1422136

Sweller, J., Mawer, R. F. & Ward, M. (1983). Development of expertise in mathematical problem solving. *Journal of Experimental Psychology: General*, *112*, 639–661.

Sweller, J., van Merriënboer, J. J. G. & Paas, F. (2019). Cognitive Architecture and Instructional Design: 20 Years Later. *Educational Psychology Review*, *31* (2), 261–292. doi: 10.1007/s10648-019-09465-5

Sweller, J., van Merriënboer, J. J. G. & Paas, F. G. W. C. (1998). Cognitive Architecture and Instructional Design. *Educational Psychology Review*, *10* (3), 251–296.

van Gog, T., Kirschner, F., Kester, L. & Paas, F. (2012). Timing and Frequency of Mental Effort Measurement: Evidence in Favour of Repeated Measures. *Applied Cognitive Psychology*, *26* (6), 833–839. doi: 10.1002/acp.2883

van Gog, T. & Paas, F. G. W. C. (2008). Instructional Efficiency: Revisiting the Original Construct in Educational Research. *Educational Psychologist*, *43* (1), 16–26.

van Merriënboer, J. J. G. & Sweller, J. (2005). Cognitive Load Theory and Complex Learning: Recent Developments and Future Directions. *Educational Psychology Review*, *17* (2), 147–177. doi: 10.1007/s10648-005-3951-0

Wierwille, W. W. & Eggemeier, F. T. (1993). Recommendations for Mental Workload Measurement in a Test and Evaluation Environment. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *35* (2), 263–281. doi: 10.1177/001872089303500205

Wirth, J. (2009). Promoting Self-Regulated Learning Through Prompts. *Zeitschrift für Pädagogische Psychologie*, *23* (2), 91–94. doi: 10.1024/1010-0652.23.2.91

Wirzberger, M., Herms, R., Esmaeili Bijarsari, S., Eibl, M. & Rey, G. D. (2018). Schema-related cognitive load influences performance, speech, and physiology in a dual-task setting: A continuous multi-measure approach. *Cognitive research: principles and implications*, *3* (1), 46. doi: 10.1186/s41235-018-0138-z

Zagermann, J., Pfeil, U. & Reiterer, H. (2016). Measuring Cognitive Load using Eye Tracking Technology in Visual Computing. In M. Sedlmair, P. Isenberg, T. Isenberg, N. Mahyar & H. Lam (Hrsg.), *Proceedings of the Beyond Time and Errors on Novel Evaluation Methods for Visualization - BELIV '16* (S. 78–85). New York, New York, USA: ACM Press. doi: 10.1145/2993901.2993908

Zander, S. (2010). *Motivationale Lernervoraussetzungen in der Cognitive Load Theory.* Berlin: Logos-Verl.

Zimbardo, P. G. & Gerrig, R. J. (2008). *Psychologie* (18., aktualisierte Auflage Aufl.). Hallbergmoos: Pearson.

Zumbach, J. & Mohraz, M. (2008). Cognitive load in hypermedia reading comprehension: Influence of text type and linearity. *Computers in Human Behavior*, *24* (3), 875–887. doi: 10.1016/j.chb.2007.02.015

Abbildungsverzeichnis